DOCUMENT RESUME

ED 103 476 TM 004 322

AUTHOR	Shuford, Emir H., Jr.; Brown, Thomas A.
TITLE	Rationale of Computer-Administered Admissible Probability Measurement.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	Advanced Research Projects Agency (DOD), Washington, D.C.
REPORT NO	ARPA Order-189-1; ARPA-P-1371
PUB DATE	Jul 74
NOTE	81p.
EDRS PRICE	MF-$0.76 HC-$4.43 PLUS POSTAGE
DESCRIPTORS	Bias; Computer Programs; *Confidence Testing; Decision Making; Feedback; *Measurement Techniques; Multiple Choice Tests; *Probability; *Response Style (Tests); Scoring Formulas; Statistical Analysis; *Testing; Test Reliability; Test Validity; Test Wiseness; Time Factors (Learning)

ABSTRACT

A student's choice of an answer to a test question is a coarse measure of his knowledge about the subject matter of the question. Much finer measurement might be achieved if the student were asked to estimate, for each possible answer, the probability that it is the correct one. Such a procedure could yield two classes of benefits: (a) students could learn the language of numerical probability and use it to communicate uncertainty, and (b) the learning of other subjects could be facilitated. This report describes the rationale underlying a procedure for eliciting personal estimates of probabilities utilizing a proper scoring rule, and illustrates some new techniques for calibrating those probabilities and providing better feedback to students learning to assess uncertainty. In addition, new results are presented comparing the incentive for study, rehearsal, and practice provided by the proper scoring rule with that provided by the simple choice procedure, and concerning the potential effect of cutoff scores and prizes upon student behavior. (Author)

BEST COPY AVAILABLE

ARPA ORDER NO.: 189-1
3D20 Human Resources

ED103476

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

R-1371-ARPA
July 1974

Rationale of Computer-Administered Admissible Probability Measurement

Emir H. Shuford, Jr. and Thomas A. Brown

TM 004 322

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

The research described in this Report was sponsored by the Defense Advanced Research Projects Agency under contract No. DAHC15-73-C-0181. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

ARPA ORDER NO.: 189-1
3D20 Human Resources

R-1371-ARPA
July 1974

Rationale of Computer-Administered Admissible Probability Measurement

Emir H. Shuford, Jr. and Thomas A. Brown

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

Published by The Rand Corporation

PREFACE

This report was prepared as part of Rand's DoD Training and Manpower Management Program, sponsored by the Human Resources Research Office of the Defense Advanced Research Projects Agency (ARPA). With manpower issues assuming an even greater importance in defense planning and budgeting, it is the purpose of this research program to develop broad strategies and specific solutions for dealing with present and future military manpower problems. This includes the development of new research methodologies for examining broad classes of manpower problems, as well as specific problem-oriented research. In addition to providing analysis of current and future manpower issues, it is hoped that this research program will contribute to a better general understanding of the manpower problems confronting the Department of Defense.

We believe decision-theoretic psychometrics holds considerable promise for military selection, training, and other applications. In the past, use of this technique has been hampered by the need to orient people to a new way of answering questions, and the need to process the much greater amount of information the method yields.

Because computers now offer a reasonable and, in many cases, a cost-attractive solution to these problems, we have devised programs and procedures for the on-line administration of tests according to the requirements of decision-theoretic psychometrics. At this time, these programs are running on certain IBM 360/370 computer systems with graphic capability, on the IMLAC PDS-1 "smart terminal" computer, and on the PLATO IV system.

This report provides the rationale for these applications, and thus should be of interest to potential users and adapters of these programs, as well as to educators interested in examining in depth the implications of this new methodology.

BEST COPY AVAILABLE

SUMMARY

A student's choice of an answer to a test question is a coarse measure of his knowledge about the subject matter of the question. Much finer measurement might be achieved if the student were asked to estimate, for each possible answer, the probability that it is the correct one. Such a procedure could yield two classes of benefits: (a) students could learn the language of numerical probability and use it to communicate uncertainty, and (b) the learning of other subjects could be facilitated.

This report describes the rationale underlying a procedure for eliciting personal estimates of probabilities utilizing a proper scoring rule, and illustrates some new techniques for calibrating those probabilities and providing better feedback to students learning to assess uncertainty. In addition, new results are presented comparing the incentive for study, rehearsal, and practice provided by the proper scoring rule with that provided by the simple choice procedure, and concerning the potential effect of cutoff scores and prizes upon student behavior.

A companion report describes an interactive computer program incorporating these procedures. See W. L. Sibley, *A Prototype Computer Program for Interactive Computer Administered Admissible Probability Measurement*, R-1258-ARPA, April 1974.

BEST COPY AVAILABLE

CONTENTS

PREFACE iii

SUMMARY v

Section
1. ELICITATION OF PERSONAL PROBABILITIES IN EDUCATION .. 1

2. THE CONTEXT OF TESTING 3

3. KNOWLEDGE AS A PROBABILITY DISTRIBUTION 3

4. THE EFFECT OF LIMITING RESPONSE OPTIONS 5

5. STRATEGIES FOR RESPONDING TO A TEST ITEM 8
   5.1 Simple Choice Testing 8
   5.2 Confidence Testing 10
   5.3 Admissible Probability Measurement 15

6. MARSHALING FACTS AND REASONS BEFORE RESPONDING 20

7. DETECTING BIAS IN THE ASSIGNMENT OF PROBABILITIES ... 23

8. PERCEIVED VERSUS ACTUAL INFORMATION 29

9. THE CONSEQUENCES OF BIASED PROBABILITIES 37

10. A LIKELIHOOD RATIO MEASURE OF PERSPICACITY 38

11. POTENTIAL IMPACT OF TESTING METHOD UPON STUDY BEHAVIOR 42
   11.1 Allocation of Study Effort Among Topics 43
   11.2 Investment of Study Effort in a Single Topic 46

12. IMPACT OF INAPPROPRIATE REWARDS UPON TEST-TAKING BEHAVIOR 47

13. SUMMARY AND CONCLUSIONS 53

Appendix
A. FITTING A PLANAR REALISM FUNCTION 55
B. HOW TO CALCULATE THE VALUE OF "p" AT WHICH MAXIMUM EXPECTED RETURN PER UNIT EFFORT IS ACHIEVED 61
C. ALLOCATING STUDY EFFORT TO MAXIMIZE PROFIT 66
D. SOME RESULTS OF OPTIMAL STRATEGIES TO ACHIEVE A PASSING GRADE 71

REFERENCES 75

RATIONALE OF COMPUTER-ADMINISTERED ADMISSIBLE PROBABILITY MEASUREMENT

1. ELICITATION OF PERSONAL PROBABILITIES IN EDUCATION

Along with the recent growth of the theory and application of the mathematics of decisionmaking has come an increased interest in expressing uncertainty in terms of personal probabilities. Most of the attention in this area has been focused upon eliciting personal probabilities from decisionmakers and experts to guide policy decisions [1-3]. However, at the end of his comprehensive and excellent review of this area, Savage [3] refers to potential educational applications of these techniques and states:

> Proper scoring rules* hold forth promise as more sophisticated ways of administering multiple-choice tests in certain educational situations. The student is invited not merely to choose one [answer] (or possibly none) but to show in some way how his opinion is distributed over the [answers], subject to a proper scoring rule or a rough facsimile thereof.
>
> Though requiring more student time per item, these methods should result in more discrimination per item than ordinary multiple-choice tests, with a possible net gain. Also, they seem to open a wealth of opportunities for the educational experimenter.
>
> Above all, the educational advantage of training people--possibly beginning in early childhood--to assay the strengths of their own opinions and to meet risk with judgment seems inestimable. The usual tests and the language habits of our culture tend to promote confusion between certainty and belief. They encourage both the vice of acting and speaking as though we were certain when we are only fairly sure and that of acting and speaking as though the opinions we do have were worthless when they are not very strong.
>
> Effects of nonlinearity in educational testing† deserve some thought, but presumably nonlinearity is not a severe threat when a test consists of a large number of items. One source of nonlinearity that has been pointed out to me is this. A

---

*Described and discussed in Sec. 5.3.

†These effects are discussed in Sec. 12.

> student competing with others for a single prize is motivated to respond so as to maximize the probability that his score will be the highest of all. This need not be consistent with maximizing his expected score, and presumably situations could be devised in which the difference would be important.

This brief statement characterizes both the promises and the problems of eliciting personal probabilities from students. The promises come from two educational goals that might be served by this application:

1. As a subject matter and skill that is valued in and of itself. For example, it is important for students to learn to discriminate degrees of uncertainty and to be able to communicate uncertainty using the language of numerical probability.
2. As a means of facilitating the learning of other subject matter, e.g., by providing more information about a student's state of knowledge.

The problems reside largely in two areas:

1. Students must be taught a new way of answering questions and they must overcome bad habits and inappropriate sets induced by their prior test-taking experience.
2. Great care must be exercised in insuring that the incentive structure impacting on the student does in fact correspond to that assumed in the decision-theoretic derivation of the method, i.e., the student must be motivated to attempt to maximize his expected score, rather than maximize the probability of exceeding some standard or surpassing his classmates. This is a subtle point we discuss at greater length in Sec. 12 below.

The purpose of this report is to describe the rationale underlying a procedure for eliciting personal estimates of probabilities utilizing a proper scoring rule, and to illustrate some new techniques for calibrating personal probabilities and providing better feedback to

students learning to assess uncertainty. In addition, new results are presented comparing the incentive for study, rehearsal, and practice provided by the proper scoring rule with that provided by the simple choice procedure, and concerning the potential effect of cutoff scores upon student behavior.

A companion report [4] describes an experimental version of an interactive computer program incorporating these procedures and focuses upon the first problem mentioned above.

2. THE CONTEXT OF TESTING

Students are asked a series of questions to ascertain their knowledge of the subject matter represented by the questions. A test item is composed of a question and a list of k (k = 2, 3, ...) possible answers, one and only one of which is correct. A "test" is composed of n of these items, usually answered in sequence, and where n typically has a value between 10 and 100.

3. KNOWLEDGE AS A PROBABILITY DISTRIBUTION

While a person holding the answer key is not at all uncertain about which answer to a question is designated "correct," a student may encounter a certain amount of uncertainty. In information-theoretic terms [5], that amount is

$$U = - \sum_{i=1}^{k} p_i \log_2 p_i \ ,$$

where p_i is the likelihood (according to the student's view of the situation) of the event, "Answer i is the correct answer." Because the p_i's may be viewed as probabilities of mutually exclusive and collectively exhaustive events, we have

$$0 \leq p_i \leq 1 \qquad \text{and} \qquad \sum_{i=1}^{k} p_i = 1 \ .$$

The uncertainty measure, called "entropy" by information theorists, achieves its maximum value ($\log_2 k$) when all the p_i's are equal and achieves its minimum value (zero) when one p_i is unity and the rest are zero.

There may be several sources of this uncertainty. Some examples are: the student may not be familiar with the standards and values of the writer of the test item; the student may not comprehend all of the language used in the test item; most important, he may not know enough facts and reasons to arrive unequivocally at the correct answer.

The uncertainty measure itself is unsatisfactory as a measure of useful knowledge, because it is symmetric or nondirectional with respect to the answers. According to this measure, a student would have minimal uncertainty (and maximal information) whenever one of the p_i's equals one. A student holding a probability of one for an *incorrect* answer possesses just as much information (in his own view) as does another student holding a probability of one for the *correct* answer. Uncertainty can serve as a measure of learning, but education and training is concerned with *what* is learned and must focus on the probability associated with the correct answer. Before a student is exposed to a subject matter and tries to learn it, he might be expected to be uncertain about answers to questions. If a question has three answers, the student's probability associated with the correct answer might fluctuate over time but remain close to the value of 1/3 corresponding to maximal uncertainty, as shown by the first segment of the curve in Fig. 1.

When the student begins to take an active interest in learning the subject matter, the probability might be expected to rise and begin to approach one as the student achieves greater and greater mastery of the subject matter. The student's probability associated with the correct answer when measured over time might trace a path similar to the learning curve shown in Fig. 1. Upon completion of the learning phase and if the student's knowledge or skill is not reinforced, the probability might begin to decline toward 1/3 and trace a forgetting curve such as that shown in Fig. 1.

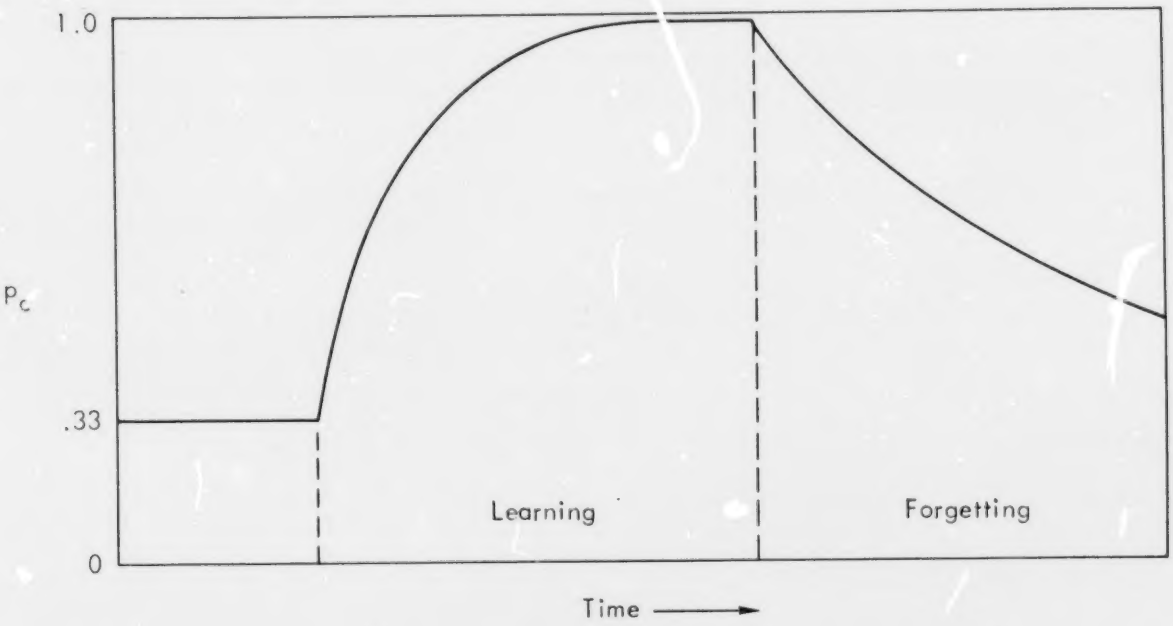


Fig. 1 — Hypothetical trace of probability over time

While these hypothetical curves resemble those found in the psychology of learning, it should be remembered that much of the experimental data in this area are reported in terms of averages over either subjects, trials, or both. Such indirect measures must be used because of the discrete nature of the responses made available to the subjects.* If it were possible to take direct and repeated measurements of a subject's personal probabilities, the need for aggregation of data would be greatly reduced and the results of experiments might appear quite different.

4. THE EFFECT OF LIMITING RESPONSE OPTIONS

In the true-false and multiple-choice methods of test administration, a student is required to select one and only one of the answers

*The major exception, response latency, is a measure continuous in the time dimension. Even so, it is frequently averaged because of its instability and, while possibly reflecting uncertainty, it fails to convey the directional information contained in the distribution of personal probabilities.

to the test question. Thus, for true-false and two-alternative multiple-choice items, the student's response is constrained to only two possible values; for three-alternative multiple-choice items, the student's response is constrained to only three possible values; and so on. If the student's state of knowledge and degree of uncertainty with respect to the question actually can assume more than k different values, it is clearly impossible to have each different response uniquely associated with a state of knowledge. The student would have to use the same response for several different states of knowledge and the restricted response set of the choice method would act as a filter inserted in the communication channel between student and teacher or experimenter. The observer of the test behavior could not use the student's response to recover unequivocably the state of knowledge that led to the response.

This limitation can be removed only by increasing the number of response options available to the student. To eliminate the filtering action described above, the number of response options must be greater than or equal to the different states of knowledge the student may possess. Because different students and the same student at different times may experience a varying number of states of knowledge and because these numbers are unknown, the safest way of preventing filtering appears to be to allow a very large number of response options.

A mathematically and graphically convenient way of doing this is to allow the student to assign a weight from the real number system to each of the possible answers to the test question. For reasons which will become apparent, let the student's response be a vector $R = (r_1, r_2, \ldots, r_k)$ where

$$0 \leq r_i \leq 1 , \qquad \sum_{i=1}^{k} r_i = 1 , \qquad \text{and} \qquad k \geq 2 .$$

Thus, for two-answer questions the student's response corresponds to selecting a point on the line segment [0,1] while for three-answer questions the response corresponds to selecting a point in an equilateral triangle as shown in Fig. 2. Questions with four possible

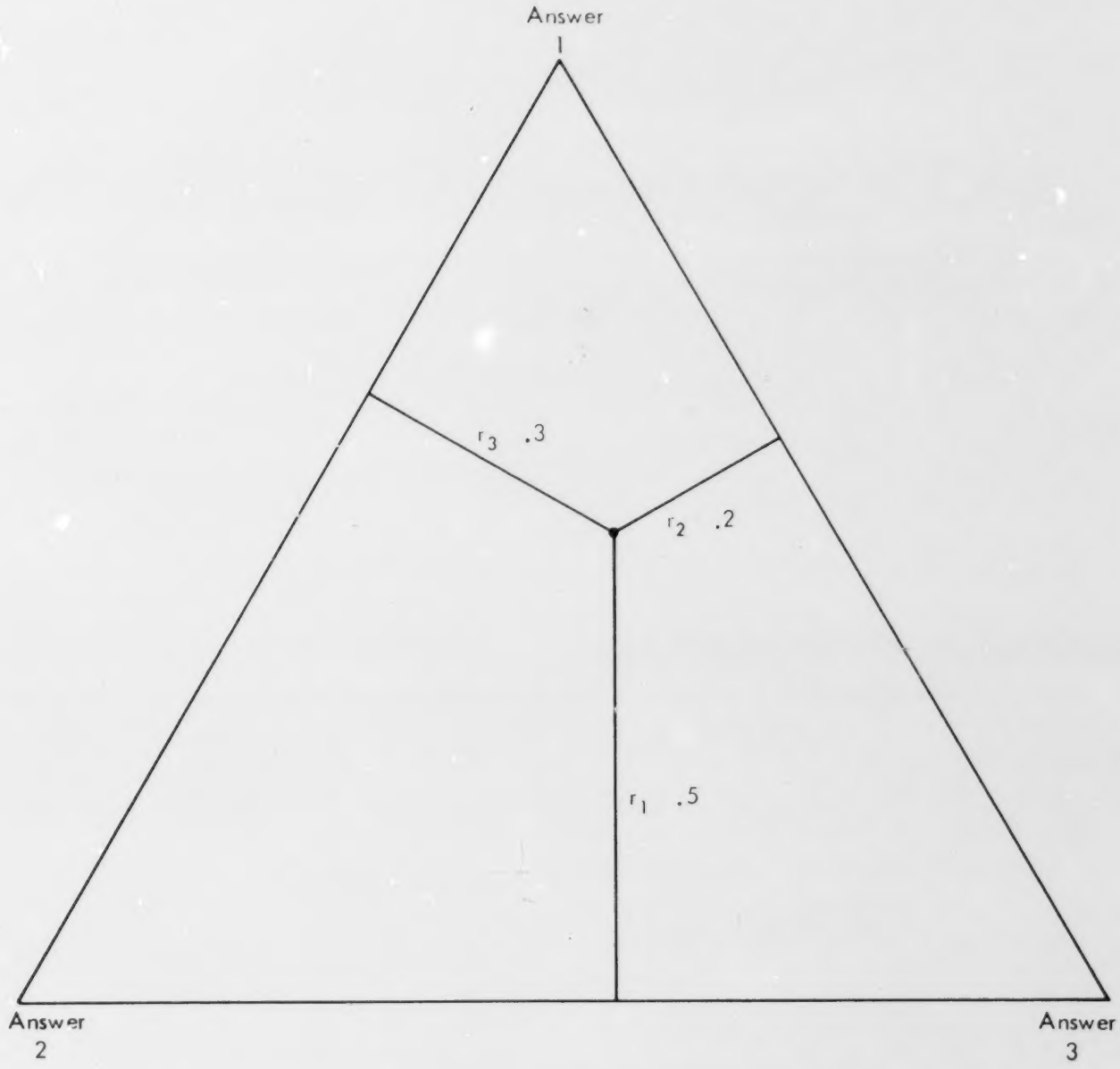


Fig. 2 — The equilateral response triangle used for computer assisted admissible probability measurement

BEST COPY AVAILABLE

answers require three dimensions for a unitary response while questions with even more possible answers require complex sequential allocation responses.

5. STRATEGIES FOR RESPONDING TO A TEST ITEM

Merely allowing a student more response options does not insure that more information about his states of knowledge will actually be transmitted. The student might, for example, exercise only a minimal number of the options or, for another example, the way he associates the response option with his probabilities might be inconsistent or arbitrary. In either event, the amount of information actually transmitted may be greatly reduced.

A student's state of knowledge, i.e., the facts recalled, reasoning, and other thought processes leading to a probability distribution over the possible answers, are directly observable only by the student himself. The student's responses are, of course, directly observable by others, but there is no biological law that a student's responses must reflect his probabilities. It is, in other words, a matter of free will and volition on the part of the student as to how he associates his response with his probabilities.

In a situation such as this, the best that can be done is to structure the task given the student so that he is rewarded for consistently and accurately associating response with his probabilities. Although the association is one-to-many, this is implicitly done with the simple choice method of responding used in the administration of achievement, aptitude, and ability tests.

5.1 Simple Choice Testing

To see this, suppose a student wishes to maximize his expected test score. With the most frequently used simple choice scoring system, he earns one point for each correct answer selected and no points for an incorrect answer.* Because his test score is simply the sum of his item

*It can be assumed without loss of generality that the student receives no points if he omits an item. Thus, the loss of a fraction of a point as illustrated by use of the "correction for guessing" formula

BEST COPY AVAILABLE

scores, his expected test score can be maximized by maximizing the expectation for each item score. Thus, for any item on the test, the decision problem faced by the student is as shown in Table 1; and his optimal strategy is to choose that course of action or response associated with the maximum expected score as defined by the information available to him at the time of making the decision. This information should be reflected in the personal probability distributions as defined in Sec. 3.

Table 1

DECISION PROBLEM FACED BY STUDENT ANSWERING ITEM UNDER SIMPLE CHOICE METHOD

	Probability (That Answer May Be Correct)						
	p_1	p_2	.	.	.	p_k	
	Correct Answer						Expected
Response	1	2	.	.	.	k	Score
Choose answer 1	1	0	.	.	.	0	p_1
Choose answer 2	0	1	.	.	.	0	p_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Choose answer k	0	0	.	.	.	.	p_k

It should be remembered that the probabilities characterize the student--not the item question and answers. One answer is correct--the others are incorrect. Two different students, or the same student at different times, may very well possess different probability distributions over the answers to the same question. The probabilities reflect the information available to the student at the time he must make his decision, and provide his only guide to action.

---

does not change the structure of the task. The structure is changed, however, if the penalty for selecting a wrong answer is greater than k - 1, where k is the number of possible answers to the test question.

For a student who wishes to score well on a simple choice test, the optimal test-taking decision rule is, for each item, to select that answer he considers most likely to be correct. If two or more answers are tied for maximum probability, it makes no difference which he selects, because the expected score is the same. This decision rule may be displayed graphically for two and three possible answers as shown in Fig. 3.

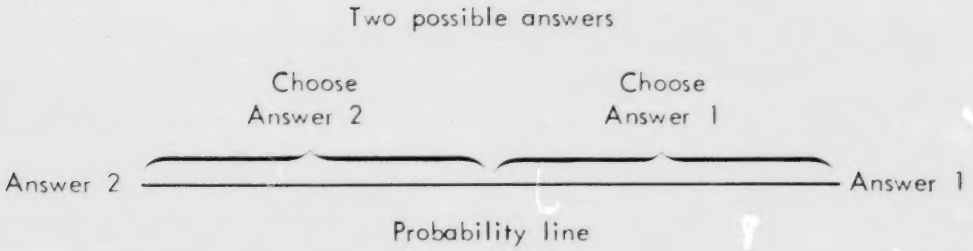
This analysis makes it apparent that while the simple choice procedure can motivate a student to use a consistent and logical mapping of probabilities onto responses, each response represents a very broad range of probabilities. When a student chooses an answer, all that may be inferred from this response is that he views no other answer as being more likely to be correct.

Terms such as "well informed," "misinformed," and "uninformed" are sometimes used to describe a person's knowledge with respect to some subject. These and related terms can be used to characterize regions of the personal probability space, as illustrated by the decomposition shown in Fig. 4 for three possible answers. Because each point on the triangle corresponds to a possible probability distribution over the three answers, this classification groups distributions that may have a similar import. For example, if a student had no reason for very strongly preferring any answer over the others, his probability distribution would be located near the center of the triangle and he would be "uninformed" with respect to the item. Figures 3 and 4 may be compared to see what information is yielded by the response-to-probability mapping induced by the simple choice method. The relations can be summarized as in Table 2.

While the simple choice response is clearly incapable of discriminating many states of knowledge, a free response such as that described in Sec. 4 would have the potential of transmitting a great deal more information about a student's state of knowledge. Will this information actually be transmitted?

5.2 Confidence Testing

Suppose, as before, that the student wishes to maximize his expected test score and that he is allowed to distribute 100 points over



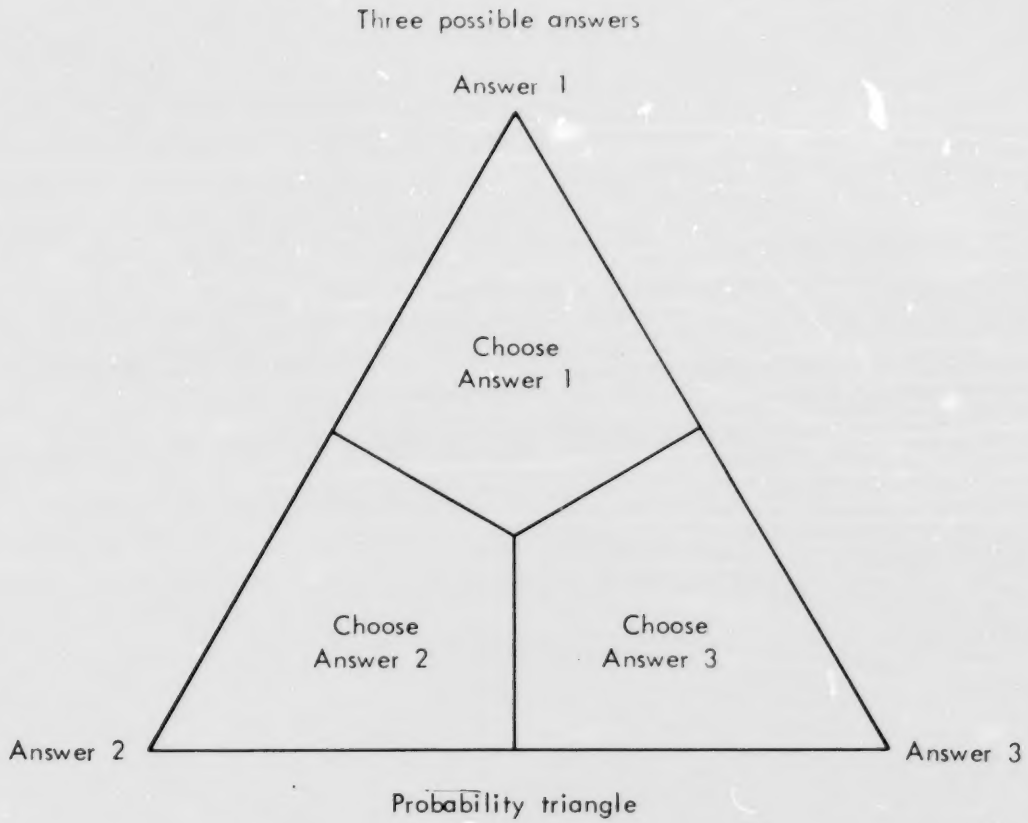


Fig. 3— Optimal decision rules for two and three answers

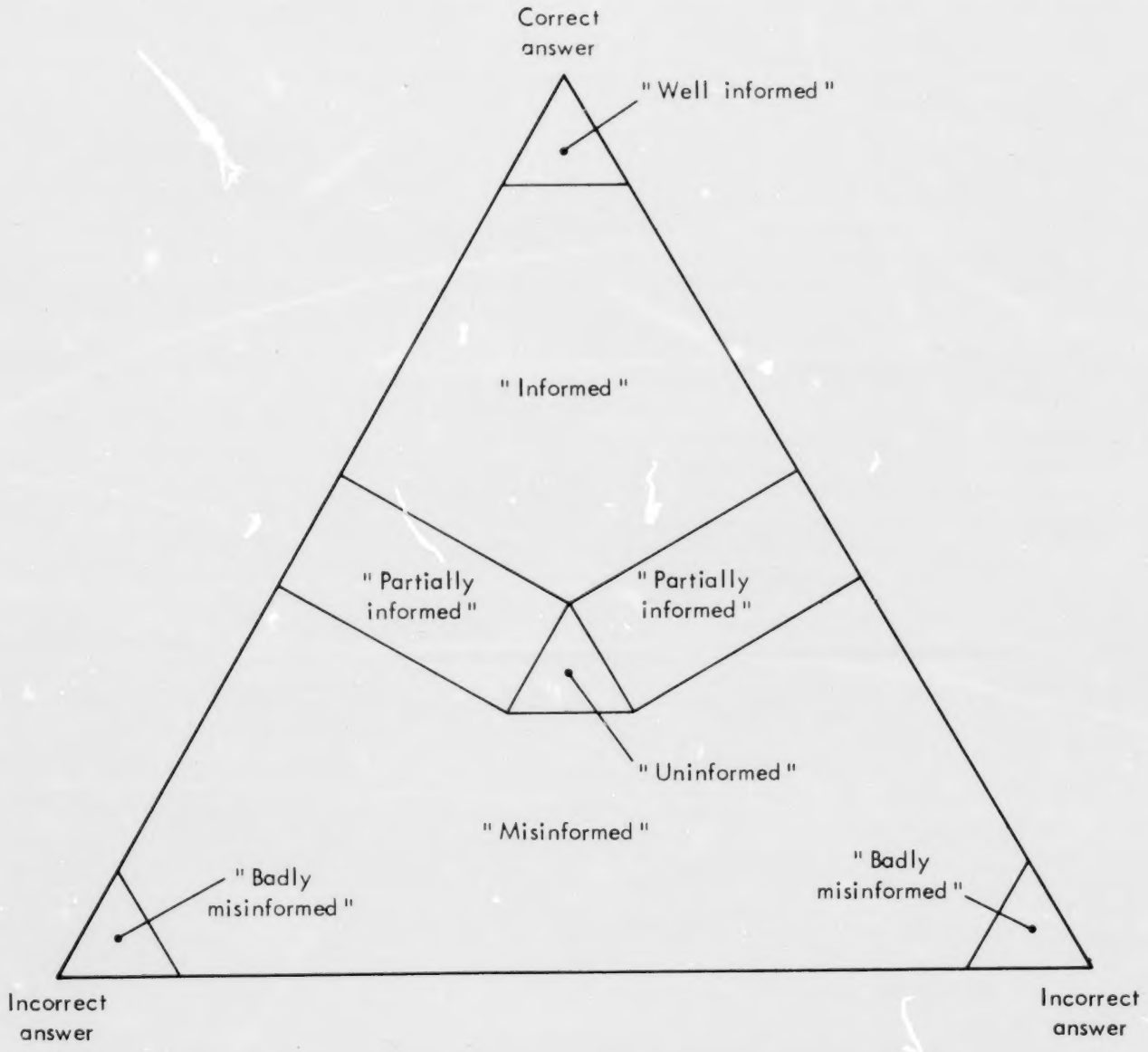


Fig. 4 — One possible decomposition of the probability triangle to represent some meaningful categories of knowledge

Table 2

INFERENCES THAT MAY BE DRAWN FROM THE SIMPLE CHOICE RESPONSE

If the student has selected	Student may be:	Student is not:
The correct answer	Well informed Informed Partially informed Uninformed	Misinformed Badly informed
An incorrect answer	Partially informed Uninformed Misinformed Badly informed	Informed Well informed

the possible answers to each item as is sometimes done in "confidence testing." This would provide a set of responses fine-grained enough to transmit considerably more information and it would be quite simple to score the student according to the number of points he allocated to the correct answer. To be more explicit, let m_{ij} be the number of points allocated on the j*th* item to the i*th* answer, where

$$0 \leq m_{ij} \leq 100 \qquad \text{and} \qquad \sum_{i=1}^{k} m_{ij} = 100 \ .$$

Let the test score be

$$\sum_{j=1}^{n} m_{*j} \ ,$$

where m_{*j} is the number of points allocated to the correct answer to item j.

The potential impact of this scoring rule upon student behavior may be investigated by finding, as before, the optimal test-taking strategy for a student who wishes to maximize his expected test score. Because his test score is simply the sum of his item scores, his expected test score can be maximized by maximizing the expected score

for each item. There are now far too many response options to list in a table, but the expected score for any allocation on a single question $(m_1, m_2, \ldots, m_k)$ may be written as

$$E(m_1, m_2, \ldots, m_k | p_1, p_2, \ldots, p_k) = m_1p_1 + m_2p_2 + \ldots + m_kp_k .$$

It is not too difficult to find the optimal decision rule, i.e., to specify for each probability distribution that response (allocation of points) which maximizes the expected item score. Because the labeling of the answers is, in a sense, arbitrary, we may assume without loss of generality that

$$p_1 \geq p_2 \geq p_3 \geq \ldots \geq p_k ,$$

i.e., the answers can be reordered from most likely to least likely to be correct in the view of the student. The decision problem is one of allocating points so as to maximize the sum of products as shown below.

$$m_1p_i + m_2p_2 + \ldots + m_kp_r .$$

The points may be placed one at a time because the placing of a point does not change the structure of the problem. Allocating a point to answer i yields a return of p_i because only that proportion p_i of the point will be added to the sum. If $p_1 > p_2$, then the first point should be placed in the first position in order to yield the largest possible return, p_i; the second point should also be placed in the first position; and so on for all 100 points. If $p_1 = p_2$, or if $p_1 = p_2 = p_3$, and so on, the points can be distributed between these maximum probabilities, but there is nothing to be gained by so doing. The optimal test-taking strategy for this scoring rule can be summarized as, "Find an answer that is at least as likely to be correct as any other and allocate all 100 points to this answer."

Thus, this scoring rule induces a mapping of response onto probability that degenerates into the simple choice situation (see Fig. 3).

BEST COPY AVAILABLE

Although many response options are offered to the student, he is maximally rewarded for placing all 100 points on the most likely answer. If a student follows this best test-taking strategy, his responses will be essentially indistinguishable from choice type responses and no additional information will be transmitted about his states of knowledge. This example shows that merely offering an increased number of response options does not guarantee that more information will be transmitted.

5.3 Admissible Probability Measurement

What is required is a scoring rule that can motivate the student to use more of the response options, each associated with a small region of probabilities. In the limit, this relation could be expressed as $r_i = f(p_i)$, where f is a monotone increasing function of p and all of the potentially available information could be transmitted. There are other cogent reasons, however, for further constraining f to be the identity function, i.e., $r_i = p_i$.

With the identity relation, the student's responses are directly interpretable as probabilities and these numerical quantities can be used in the equations of probability, information, and decision theory [1]. Students would be learning to communicate degrees of uncertainty in a universal language of probabilities. For this reason and in the absence of any compelling reasons to do otherwise, it seems reasonable to require that scoring rules possess the property that a student can maximize his expected score *if and only if* his responses match his probabilities. Scoring rules satisfying this condition have variously been called "proper" [3], "reproducing" [1,6], and "admissible" [6].

It has been shown that there exist an infinite number of scoring rules that induce the identity relation between response and probability [1,6]. Only one, however, possesses the property that the score depends only upon the response assigned to the correct answer, and not upon how the responses are distributed over the other answers when more than two answers are possible [6]. This is the logarithmic scoring rule, which may be written as $S_i = A \log r_i + B$, where $A > 0$.

Notice that $\log r_i = -\infty$ when $r_i = 0$. This means that the logarithmic scoring rule cannot be strictly applied in practice, because if a student ever assigned a response of zero to a correct answer the logarithmic scoring rule calls for an infinite penalty. However, by restricting the range of possible responses that a student may use, so that $r_i \geq d$ (where d might be set at 0.01 or some other small value) the need for a very large penalty is avoided, but with the sacrifice of some accuracy in measuring very small probabilities [6].*

For many purposes it seems desirable to adjust A and B so that when a student possesses no information with respect to an item (i.e., all p's are equal), his score will be zero. This may be accomplished by choosing a range, K, of possible scores and setting $A = 0.5K$ and $B = 0.5K \log k$, where, as before, k is the number of possible answers. The score that the student will receive if answer i is correct can now be written:

$$S(r_i) = 0.5K \log kr_i , \qquad r_i \geq 0.01 .$$

A range of 100 points appears satisfactory for many applications. Figure 5 shows graphically the conditional scores for the case of two possible answers while Fig. 6 shows selected conditonal score triplets for the case of three possible answers. Notice that in the case of two alternatives, the maximum score obtainable is about 15, while the minimum score is about -85. In the case of three alternatives, the maximum is about 44 and the minimum is about -76. This difference in maximum and minimum scores is caused by the requirement that the scoring function be zero when all the responses are equal, but it may also be taken to indicate that prediction may be, in some sense, easier with two alternatives than with three.

Notice, also, how the penalties tend to be larger than the rewards. This is a characteristic of all the admissible scoring rules because the nonlinearity is required in order to induce matching behavior in

*For those special situations requiring the accurate measurement of very small probabilities, d may be set at a very much smaller value.

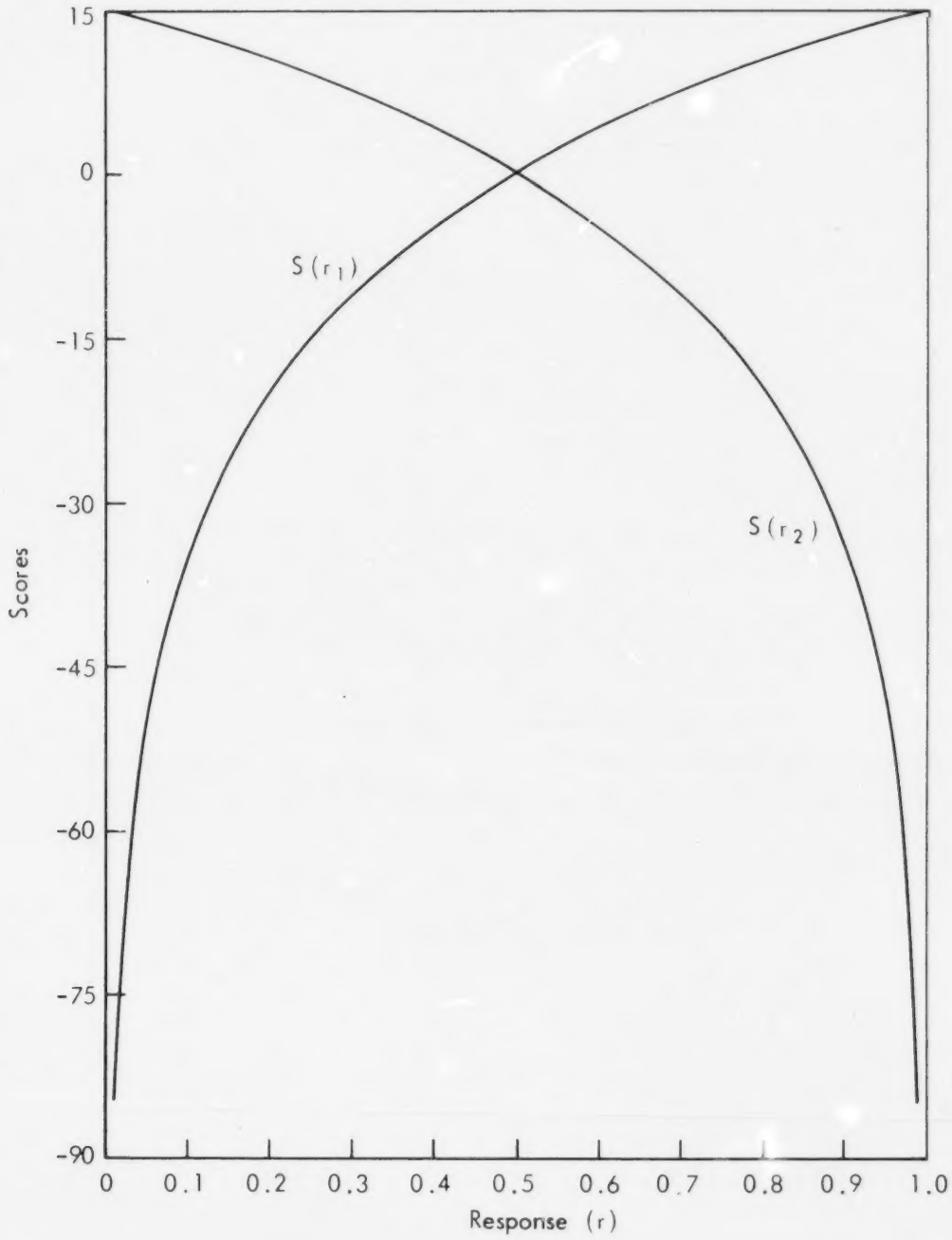


Fig. 5 — Conditional score functions for the case of two possible answers Because $r_2 = 1 - r_1$, conditional score pairs may be found where $r_1 = r$.

Fig. 6 — Conditional score triplets (based on logarithmic scoring function) for some selected responses on the equilateral probability triangle

BEST COPY AVAILABLE

the students. This characteristic of the scoring rule may have other implications, as illustrated by this quotation from a Rand staff member after experiencing computer-assisted admissible probability measurement as reported in [4].

> One thought that occurred to me after I took (the) test was that, contrary to other tests, this one can also be a *learning* experience. The situation in which one is punished severely for emphatically stating what turns out to be wrong, more so than one is rewarded for what is right even if emphatically stated, is one that is closer to the reality situation of everyday life than the simple tests that look only for right or wrong. Thus, the test itself exercises a certain negative reinforcement against stating too strongly what one is not really sure about, and thus actually conditions a person to using what knowledge he has, and at the varying degree of certainty with which he commands it, *judiciously*. This will be of advantage to him in life. For it is a fact of life that a mistake stated with aplomb permanently reduces our credibility with others who must rely on our say-so, i.e., it makes us less likely to succeed in a job, for instance. Thus (the) test is not only evaluative but educational.

Consider now the optimal test-taking strategy for a student who wishes to maximize expected test scores. As before, the total test score is simply the sum of the item scores, so expected test scores can be maximized by maximizing each expected item score, which may be expressed as

$$E[S(r_1), S(r_2), \ldots, S(r_k) | p_1, p_2, \ldots, p_k]$$

$$= E[S(r)|p] = \sum_{i-1}^{k} p_i S(r_i)$$

$$= \sum_{i=1}^{k} p_i (0.5K \log kr_i)$$

$$= 0.5K(\log k + \sum_{i=1}^{k} p_i \log r_i) .$$

This last form of the equation makes clear a relation between the logarithmic scoring rule and information theory. If a student responds with $r_i = p_i$ for all i, then his expected item score is proportional to a constant plus the amount of information he perceives that he possesses with respect to the item. This relation makes it easy to derive information measures from test scores based on the logarithmic scoring rule (cf. Sec. 8).

Should the student respond with his probabilities or, more specifically, how does the logarithmic scoring rule induce the student to do this in order to maximize his expected scores? Figure 7 shows, for the case of two answers, expected scores for all possible responses for each of the four different probability distributions, while Fig. 8 shows expected score contours for the case of three answers. Notice that for each probability distribution the largest expected score occurs where the response matches the probability distribution. With an admissible scoring rule such as the logarithmic, this is true not only for these selected distributions but for all possible probability distributions. This means that a student always suffers a loss in expected score whenever he deviates from the optimal test-taking strategy of setting $r_i = p_i$ for all i. Note further that the loss in expected score increases the more he deviates from this optimal strategy. For those instances in which the student has no knowledge about an item, i.e., all the p's are equal, if he pretends to have complete knowledge by setting one of the $r_i = 1$, he loses 35 points in expected score when there are two answers and almost 43 points when there are three answers. This feature of the logarithmic scoring rule may be expected to serve as a disincentive toward guessing-type behavior. More important, however, the logarithmic scoring rule can serve to induce an exact association of responses with probabilities. What other impact might it have upon student behavior?

6. MARSHALING FACTS AND REASONS BEFORE RESPONDING

Up to this point the decision analyses have taken the student's uncertainty (his probability distribution) as given, and then focused on finding that response which gives the highest possible expected

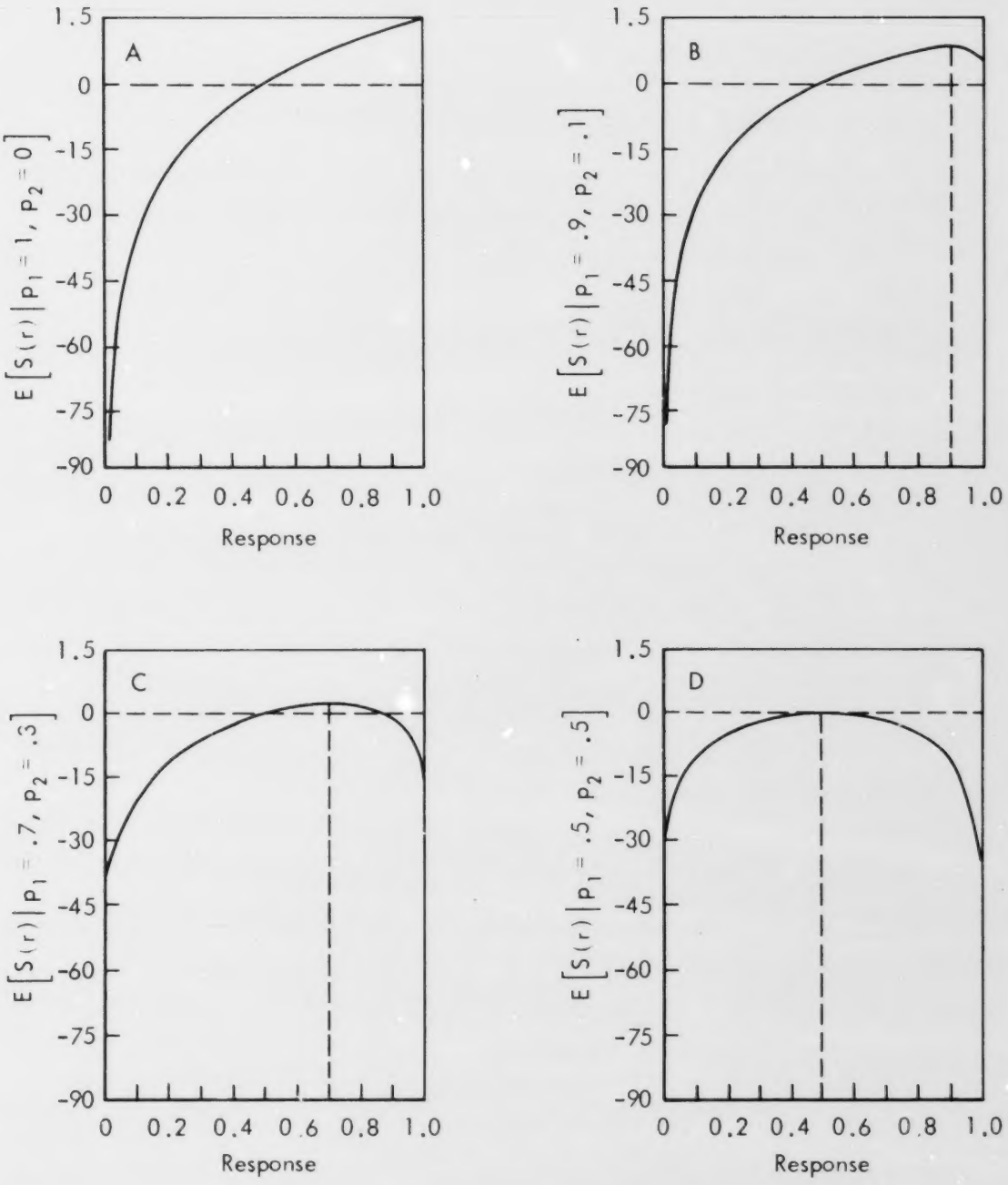
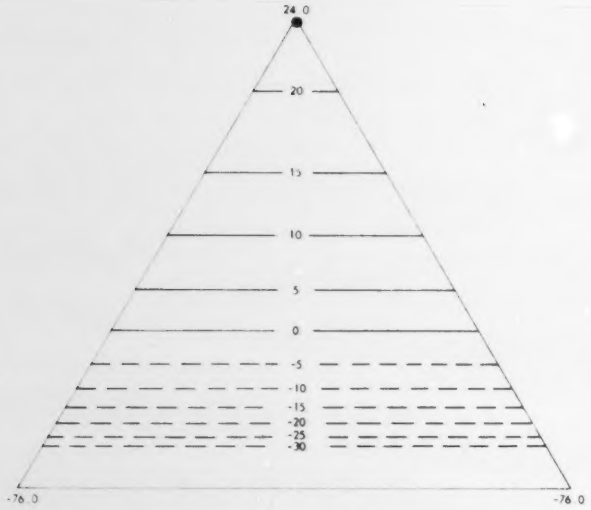
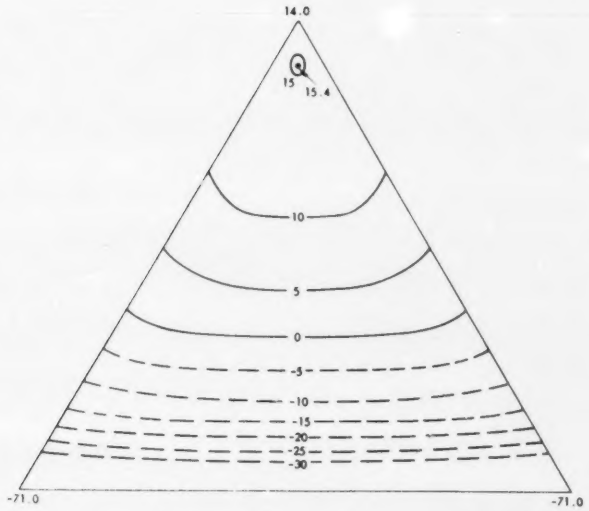


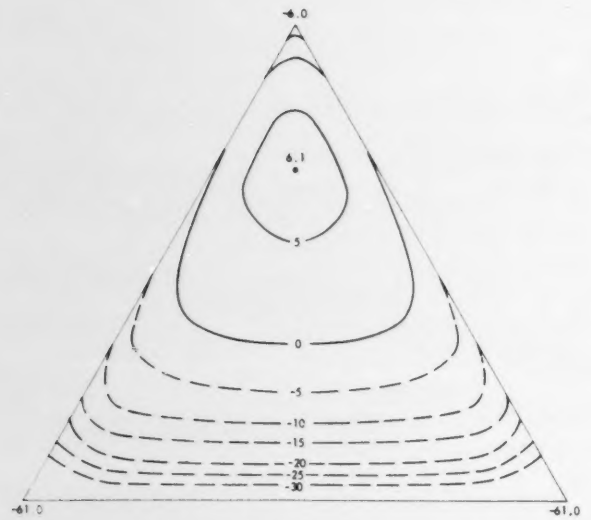
Fig. 7— Expected score functions in the case of two answers for four selected probability distributions



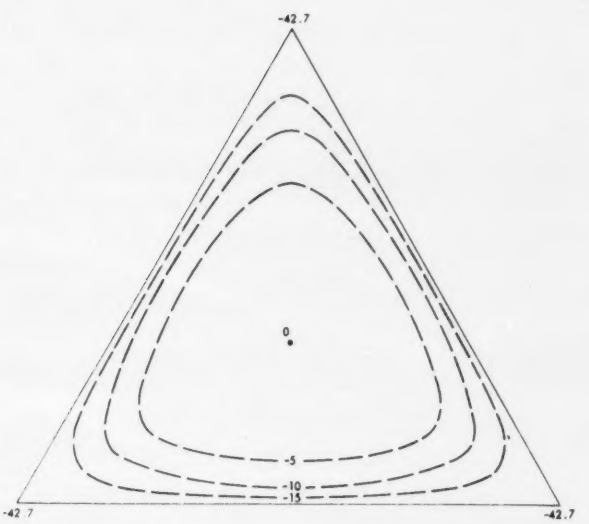
$E\left[S(r)|p_1=1,\ p_2=0,\ p_3=0\right]$



$E\left[S(r)|p_1=.9,\ p_2=.05,\ p_3=.05\right]$



$E\left[S(r)|p_1=.7,\ p_2=.15,\ p_3=.15\right]$



$E\left[S(r)|p_1=1/3,\ p_2=1/2,\ p_3=1/3\right]$

Fig. 8 — Expected score contours in the case of three answers for four selected probability distributions

BEST COPY AVAILABLE

score. There does come a time during the taking of any test when the student has to commit himself to some response. The optimal strategies derived above are appropriate to this problem and, thus, are designated test-taking strategies in the narrow sense.

The scope of the decision context must be enlarged, however, when it is considered that a student may have some control over his probability distribution for an item. For one example, while taking a test he can think more deeply about the questions and answers to bring more facts and reasons to bear upon the problem at hand. For another example, prior to taking a test he can study in order to gain additional information about the subject matter. What implications does the scoring rule have for these types of behavior on the part of a student?

Given that a student uses the optimal response strategy, r^*, his optimal expected score, $E[S(r^*)|p] = \sum_{i=1}^{k} p_i S(r_i^*)$, can be computed for each possible probability distribution. Figure 9 shows this relation when there are two possible answers for both the simple choice or linear and the logarithmic scoring rules. Notice that as the student acquires information to move his probability away from the state of being uninformed ($p_1 = p_2 = 0.5$), the optimal expected score from the simple choice procedure increases in proportion to the distance moved along the probability scale, while that from the logarithmic scoring procedure increases only slightly at first and then more and more as higher levels of mastery are achieved. A similar effect is observed in the case of three possible answers, as shown in Figs. 10 and 11. Thus, the logarithmic procedure requires a higher level of mastery to yield any given optimal expected score (other than zero) than does the simple choice procedure and, in this sense, can serve as a more stringent incentive system for learning. In Sec. 11 we build a model to investigate this in more detail.

7. DETECTING BIAS IN THE ASSIGNMENT OF PROBABILITIES

The central theme so far has been concerned with the relation between a student's responses and his probabilities. The probabilities were taken as given and the relation (if any) between the student's probabilities and the external world was reflected indirectly in the student's actual test score.

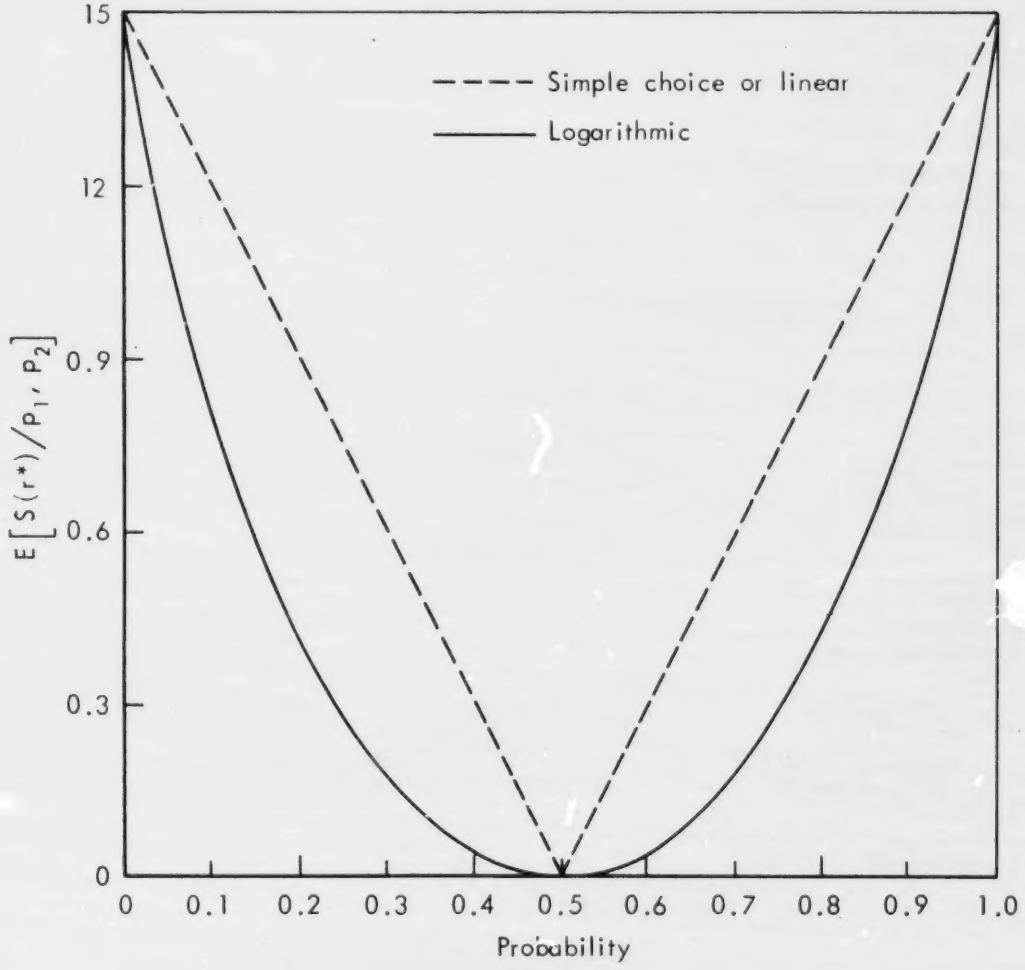


Fig. 9— Optimal expected score as a function of probability in the case of two answers

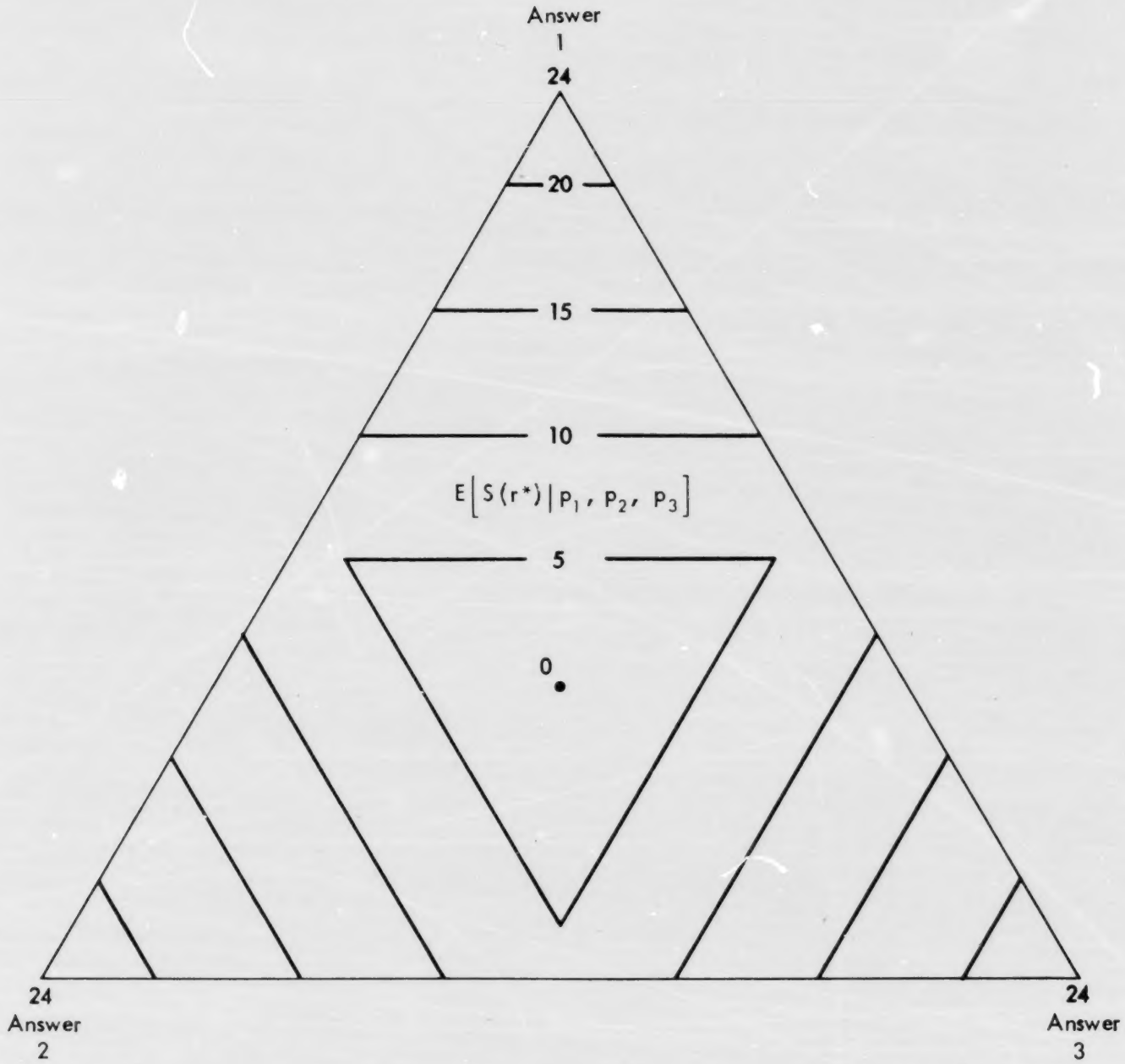


Fig. 10—Optimal expected score for simple choice and linear scoring rules in the case of three answers

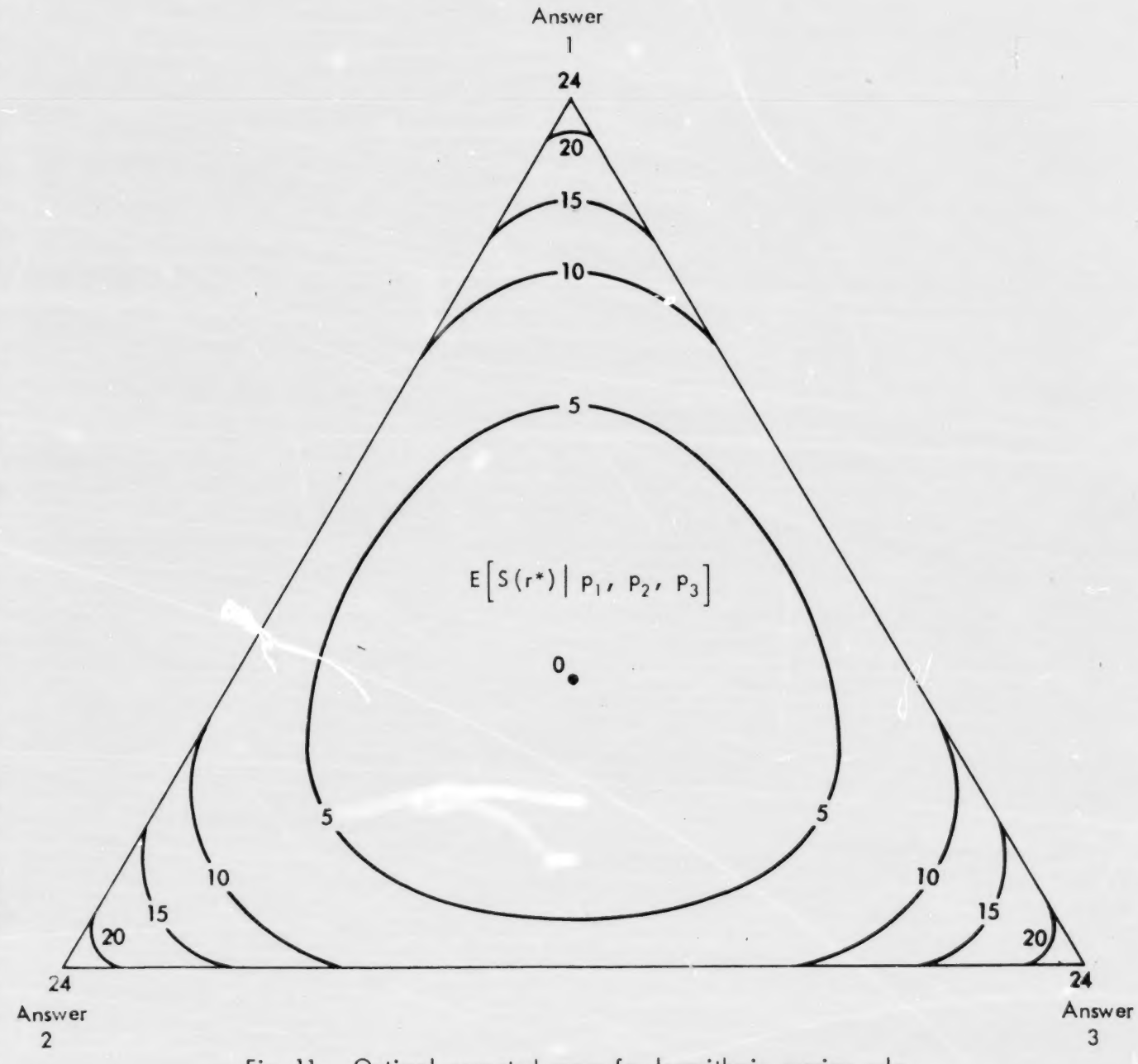


Fig. 11 — Optimal expected score for logarithmic scoring rule in the case of three answers

Here, the focus will be on the assessment of probabilities themselves, i.e., on the relation between the student's probabilities and the facts and reasons leading to these probabilities. It should be recognized that there is a point beyond which this type of analysis cannot go. There exists no completely general descriptive or prescriptive theory of how to derive probabilities from facts and reasons. Even if such a theory did exist, there is at present no way of knowing what facts and reasons a student is aware of at a particular moment in time. Nevertheless, a number of powerful methods for the assessment of probabilities are currently available or under development.

The *external validity graph* is the most fundamental means of calibrating and operationally defining personal probabilities. Assume that a student taking a test is following the optimal test-taking strategy for the logarithmic scoring rule so that $r = p$. Now let

$N(C|p)$ = number of correct answers assigned probability p ,

and

$N(I|p)$ = number of incorrect answers assigned probability p .

Then

$$r(p) = \frac{N(C|p)}{N(C|p) + N(I|p)}$$

is the empirical success ratio conditional upon the probability assignment p. A student's probability assignments are perfectly valid if $R(p) = p$ for all p when the number of observations is increased without limit. Figure 12 illustrates an external validity graph.

An external validity graph requires an inordinate amount of data before a student's probabilities can be calibrated. However, by placing some constraints on the relation between relative frequency and probability, it is possible to obtain some results with much less data. Suppose, now, that $R(p)$ tends to $q = ap + b$, $0 \leq q \leq 1$. To estimate

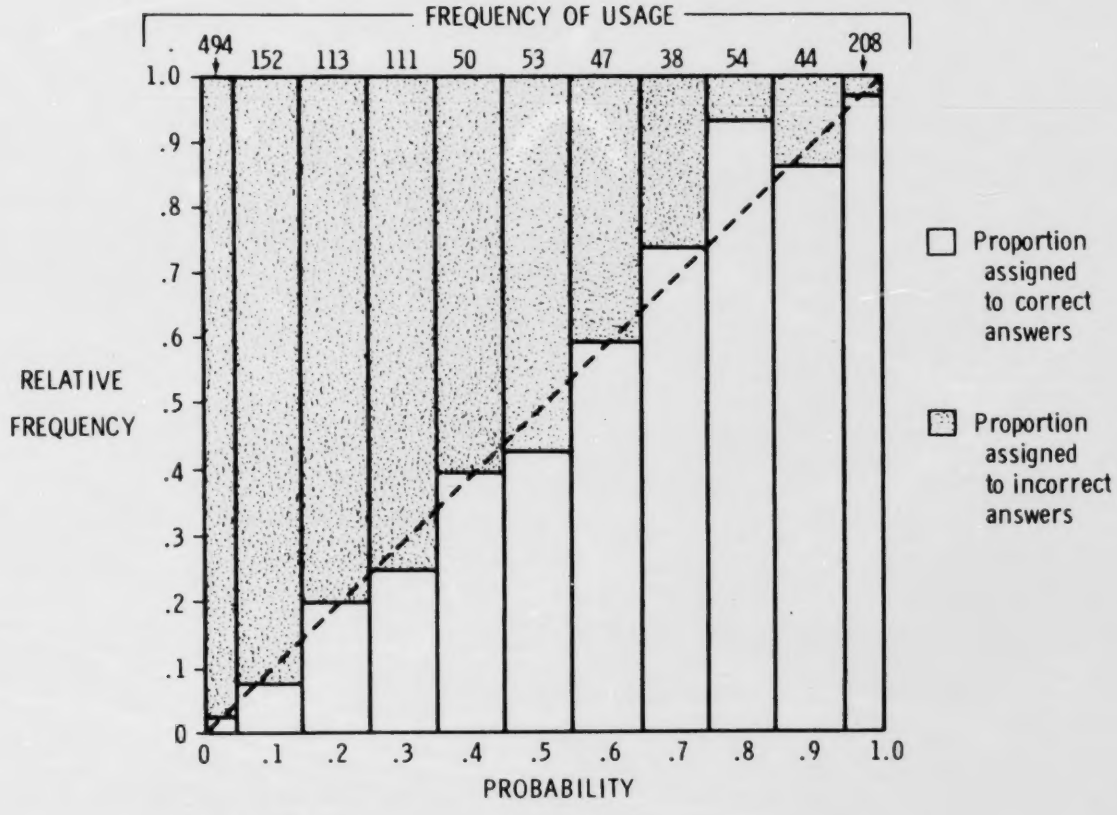


Fig. 12 — An external validity graph based on 28 15- and 20-item tests taken by one subject after receiving training in admissible probability measurement. All tests were composed of three answer items. Dashed line represents ideal match between relative frquency and probability.

this linear realism function, let $p_1 < p_2 < \cdots < p_L$ be the level of probability that the student has assigned, and let

u_i = number of times p_i has been assigned to a correct answer, and
v_i = number of times p_i has been assigned to an incorrect answer.

A convenient estimation procedure is to find a and b so as to minimize

$$\sum_{i=1}^{L} (u_i + v_i) \left(\frac{u_i}{u_i + v_i} - ap_i - b \right)^2 .$$

The least square estimators are (see [15]):

$$\hat{a} = \frac{\sum(u_i + v_i) \sum u_i p_i - \sum(u_i + v_i) p_i \sum u_i}{\sum(u_i + v_i) p_i^2 \sum(u_i + v_i) - [\sum(u_i + v_i) p_i]^2}$$

$$\hat{b} = \frac{- \sum(u_i + v_i) p_i \sum u_i p_i + \sum(u_i + v_i) p_i^2 \sum u_i}{\sum(u_i + v_i) p_i^2 \sum(u_i + v_i) - [\sum(u_i + v_i) p_i]^2} .$$

As long as a reasonably wide range of p's is used by the student, this estimation procedure can yield fairly stable results with 15- and 20-item tests, so it represents a tremendous improvement in efficiency over the external validity graph. It should be noted that if the slope estimate $\hat{a} > 1$, the student appears to be undervaluing his subject matter information, while if $\hat{a} < 1$, the student is apparently overvaluing his information (see Fig. 13). This analysis of bias appears to be completely satisfactory for the case of just two possible answers to each test item. Where three or more answers are allowed, however, this analysis requires that each response/probability is independent of the others in the distribution. This is not necessarily true for all persons. For example, some people might tend to overvalue information when deducing reasons in favor of an answer, but tend to undervalue information when deducing reasons against an answer. In Appendix A we give a planar estimation procedure for the case of three possible answers. This procedure is capable of detecting the separate dimensions of bias.

The calibration results yielded by the realism function are related not only to Savage's conjecture quoted at the beginning of this report but also to a familiar saying of Confucius: "When you know a thing, to hold that you know it and when you do not know a thing, to acknowledge that you do not know it. This is knowledge."

8. PERCEIVED VERSUS ACTUAL INFORMATION

This aspect of student behavior may be explored further by computing (under the assumption of independence among test items) the

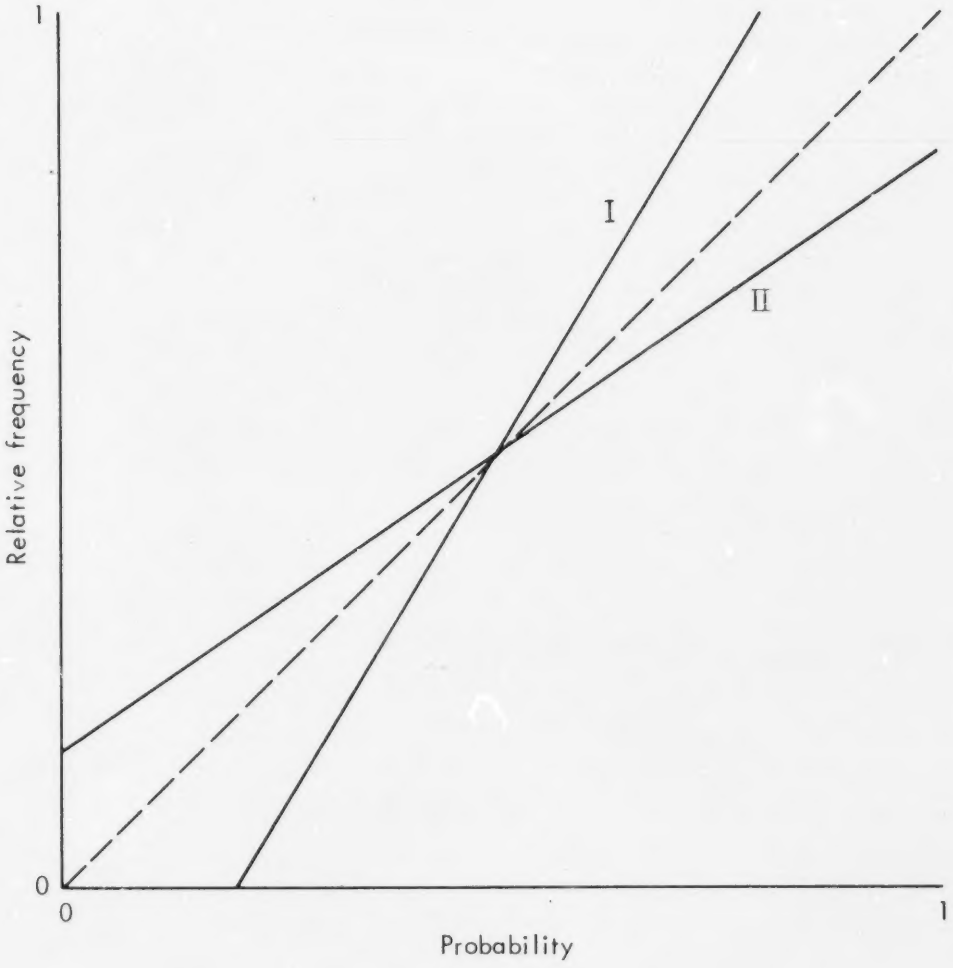


Fig. 13— Two realism functions based on probability assignments for two answer questions. Person I undervalues his information while person II overvalues his information

amount of information the student *perceives* he possesses with respect to the subject matter of the test, as indicated by his probability assignment; i.e.,

$$n \log k + \sum_{j=1}^{n} \sum_{i=1}^{k} p_{ij} \log p_{ij} .$$

If the logarithmic scoring rule is used, this expression when multiplied by 0.5K becomes the difference between the test score the student

BEST COPY AVAILABLE

expects and the test score he would expect if he had no information relevant to the subject matter.

The amount of information the student actually possesses with respect to the subject matter of the test may be estimated by substituting $\hat{p}_{ij} = \max[0, \min(1, \hat{a}p_{ij} + \hat{b})]$ for the p_{ij} in the above expression. Comparison of these two information measures reflects the extent and direction of student bias.* This comparison may be made graphically in terms of the *information square* shown in Fig. 14, which has been drawn to illustrate the aptness here of the Arabian proverb,

> He who knows and knows that he knows,
> He is wise, follow him.
>
> He who knows and knows not that he knows,
> He is asleep, awaken him.
>
> He who knows not and knows not that he knows not,
> He is a fool, shun him.
>
> He who knows not and knows that he knows not,
> He is a child, teach him.

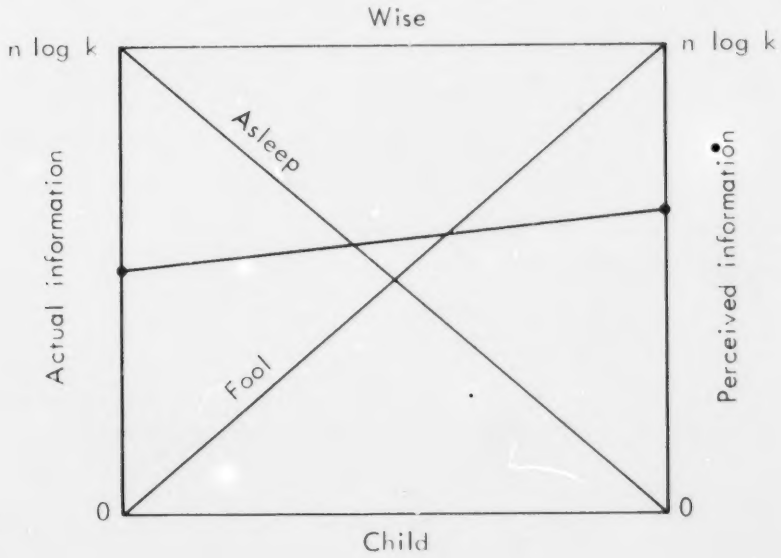


Fig. 14 - The information square

*Under certain conditions, however, the information measures may be equal but the realism function reveals that the student is tending to overvalue his information. These instances tend to be extreme and even pathological, e.g., when a student tries to minimize his expected test score.

At Rand we have demonstrated, and tried out, computer-administered decision-theoretic testing with many different people using as sample tests *Reader's Digest* vocabulary tests; Humanities, Natural Sciences, and Social Sciences items from a workbook for the College Level Examination Program tests; and a midterm postgraduate-level test in Econometrics. About halfway through these demonstrations we decided to begin keeping a permanent record of what people were doing at the terminal.

Figure 15 compares the two information measures for the first test taken by each of 66 people. Most of the data points fall below the diagonal, indicating that most of the "subjects" at least initially overvalue their knowledge of these subject matter areas. A few people fall close to the diagonal, suggesting that some people may exist who can discriminate with great accuracy what they know well from what they know less well.

What happens when people take more tests and, thus, gain more experience with decision-theoretic testing? We find that many people can reduce their score loss due to lack of realism [4]. I think that this improvement comes as they begin to experience the consequences of the admissible scoring system [6] and learn to reduce their risk-taking tendencies by making their utilities more nearly linear in points earned or lost. There does, however, appear to be a limit to this improvement.

A number of people were encouraged or challenged to take more tests, and to try to be as realistic and to score as well as they possibly could. We ended up with 11 subjects who took an appreciable number of tests--enough so we could discard the early ones they took while they were learning the procedures and the consequences of the admissible scoring system.

Figure 16 shows the apparently stable state behavior of the most biased of the 11 subjects. The line designated $\bar{I}_A$ is located at the mean of the actual information measures, while the line designated $\bar{I}_P$ is located at the mean of the perceived information measures. The intersection of the two lines gives a gross indication of actual versus perceived information for those tests the subject decided to attempt. By taking the ratio of $\bar{I}_P$ to $\bar{I}_A$ we can obtain a rough measure of the extent and direction of bias. The ratio for this subject is 2.44, indicating that she thought that she had almost two and one-half times as much information as she actually had.

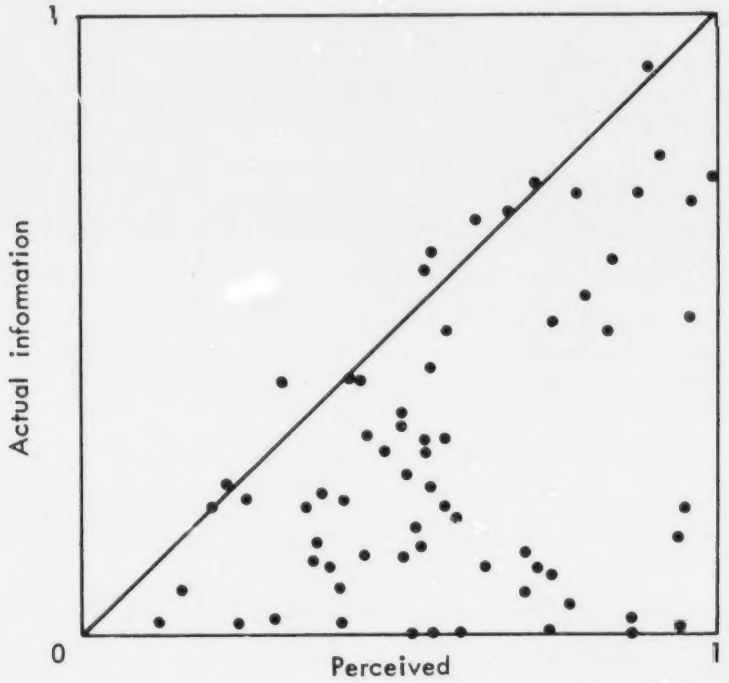


Fig. 15 — Information comparisons for 66 subjects while taking first computer-adminstered admissible probability test

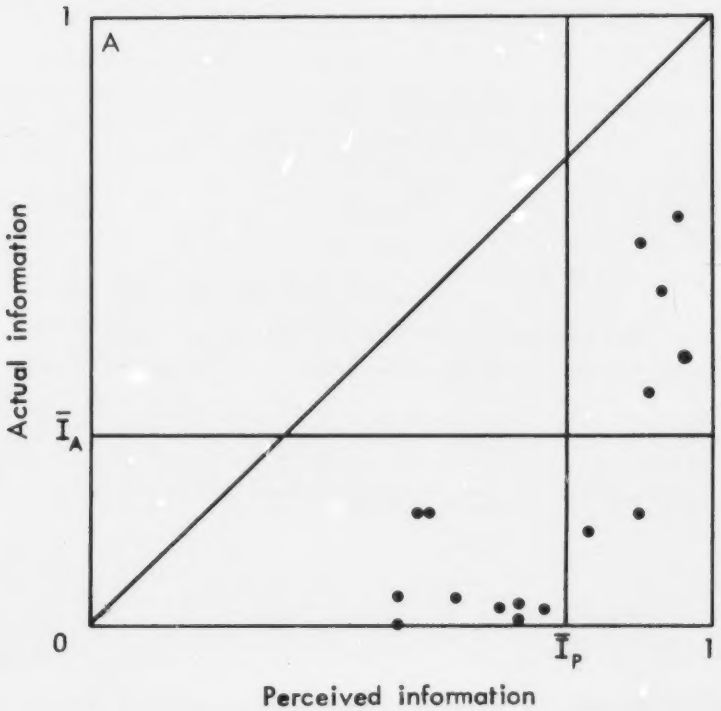


Fig. 16 — Information comparisons for subject A, the most biased subject. Early tests excluded. Data shown for last 18 tests taken by subject

Table 3 lists some personal characteristics for the 11 subjects arranged in decreasing order of bias, which goes down almost to the unbiased value of 1.00. Notice that no subject yielded an overall ratio less than one, which would have indicated a person who typically undervalued his information. Figure 17 compares the information measures for subjects B through K. Subject B, although apparently striving to reduce bias and to improve his score, was unable to do so. The remaining subjects, depicted in decreasing order of bias, were more and more often successful in producing a realistic assessment of their uncertainty. Subjects J and K, the two most accurate subjects, were remarkably consistent in demonstrating their ability to assess their uncertainties accurately.

Table 3

SUBJECT CHARACTERISTICS

Subject	$\bar{I}_P/\bar{I}_A$	$\bar{I}_A$	Tests	Sex	Age	Education
A	2.44	0.31	18	Woman	20-30	Master's +
B	2.42	0.17	12	Man	30-40	Doctorate
C	2.26	0.28	7	Man	50-60	Doctorate
D	2.11	0.32	27	Woman	20-30	Bachelor's
E	1.81	0.18	20	Woman	20-30	Some college
F	1.67	0.40	12	Woman	50-60	Doctorate
G	1.52	0.30	20	Woman	30-40	Bachelor's
H	1.33	0.35	9	Girl	9	Third grade
I	1.22	0.38	21	Girl	12	Fifth grade
J	1.02	0.71	34	Man	40-50	Doctorate
K	1.00+	0.85	8	Man	40-50	Doctorate

In conclusion, the introduction of decision-theoretic testing makes it possible to define and to measure for the first time a human ability, call it *realism*, which may prove to be a very important determinant of individual and team performance. For example, to what extent and in what manner is an unrealistic student handicapped in his attempts to learn and to study effectively? For another example, does a team of realistic people tend to out-perform a team of overvaluing people and, if so, for what types of tasks? Answers to these and many other questions must await further research.

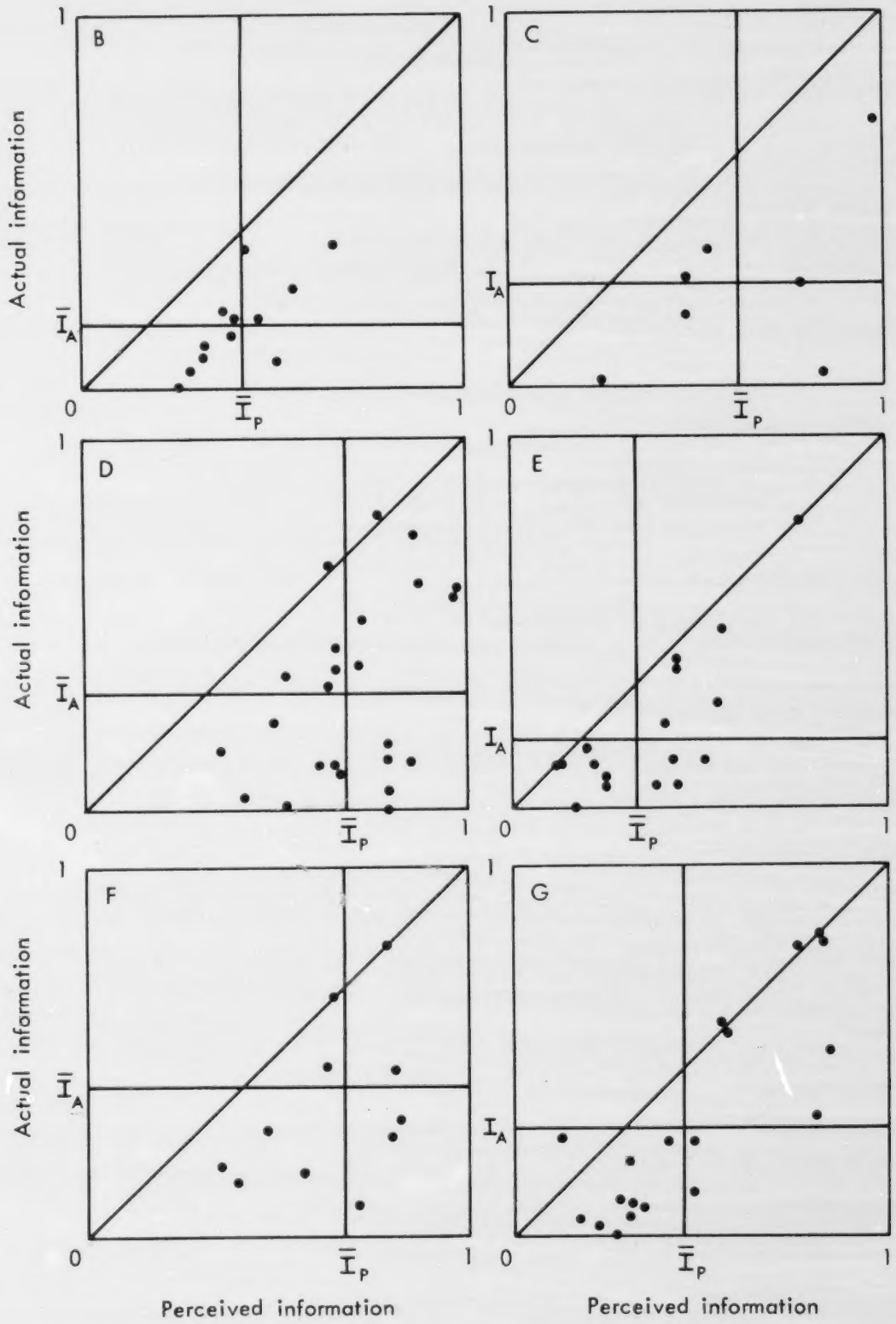


Fig. 17 — Information comparison for subjects B through K

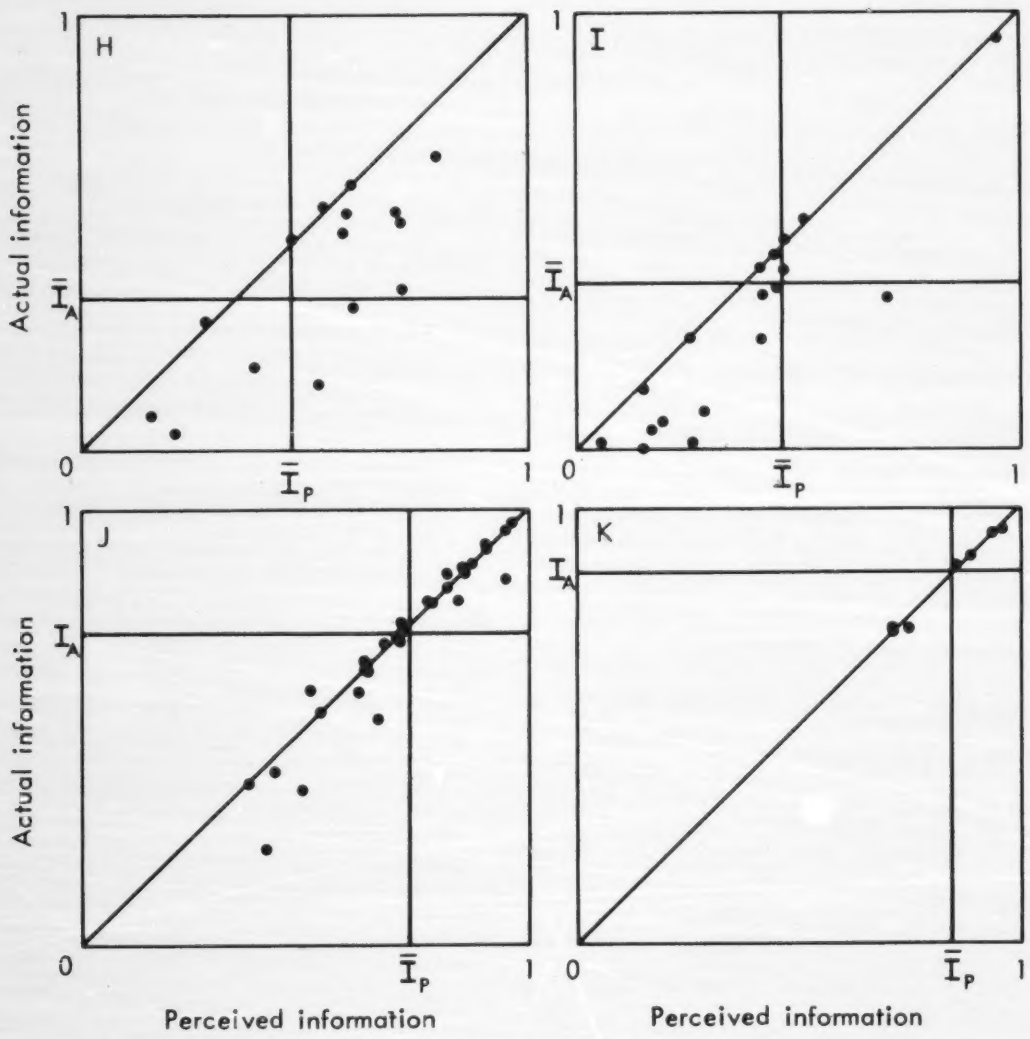


Fig. 17 — Continued

We have shown here that some people can be very realistic over a wide range of subject matter while others characteristically overvalue their information. We do not yet know what deficits in this ability exist within different subgroups of the population nor do we know exactly how to go about educating people to become more realistic. The results for subject A, summarized in Fig. 16, certainly prove that level of education does not insure realism in assessing and communicating uncertainty.

9. THE CONSEQUENCES OF BIASED PROBABILITIES

Decomposing the test score provides a convenient means for showing a student the consequences of having less than perfect realism in assessing the value of information. It is also related to a major, but little known, property of an admissible scoring system: *A student's actual test score is maximized if and only if his responses match the conditional success ratios defined in the previous section.* Thus, the effect of experience upon a student who desires to score well on admissible probability tests should be in the direction of making his responses conform more closely to the conditional success ratios.* In other words, the student should develop his ability to give better probabilistic predictions.

The maximum test score obtainable on an n-item test with the logarithmic scoring rule is $M(n) = n(0.5K \log k)$, while the minimum score is $m(n) = n(0.5K \log 0.01K)$ because of the restriction on r. If S(n) is the total test score earned by a student, then $M(n) - S(n)$ is the amount of improvement left in order to achieve perfect mastery of the test, and when $K = 100$ this total improvement score can range between 0 and 100n. Thus, one function served by the use of an improvement score is the elimination of negative scores.

This total improvement score may now be broken down into two scores, each of which has a meaningful interpretation. Suppose the test is rescored using the adjusted probabilities $\hat{p}_{ij}$, computed from the student's realism function as described above, instead of the student's actual responses r_{ij}. This procedure yields a new score, $\hat{S}(n)$, which typically is greater than or equal to S(n).† The adjusted score $\hat{S}(n)$ is an estimate of the score the student could have made if he were unbiased and

---

*For a student who is biased in assessing uncertainty, i.e., $p \neq r(p)$, we have the possibility of conflict between maximizing expected score versus maximizing actual test score. While of profound importance, a detailed treatment of this subject is beyond the scope of this report. The conflict is resolved, of course, if the student is able to change his probabilities to match the conditional success ratios.

†Recall that the realism function is only a least-squares fit to the data. If the realism function were fitted using a maximum likelihood procedure, the logarithmic score would be strictly maximized and there would be more assurance that $\hat{S}(n) \geq S(n)$.

made more effective use of the information available to him. Now, $\hat{S}(n) - S(n)$ represents the improvement possible through more effective use by the student of the information already available to him, while $M(n) - \hat{S}(n)$ represents the improvement possible as a result of his gaining additional information pertaining to the subject matter of the test. These two improvement scores are a decomposition of the total test score because, when summed, they equal the total improvement score. Such an analysis, of course, is not possible with the simple choice method.

10. A LIKELIHOOD RATIO MEASURE OF PERSPICACITY

Realism appears to be an important goal for human behavior. There is some indication, however, that it may not be sufficient as an ideal. For example, by using complex strategies which sacrifice potential test score, a student might be able to produce a realism function with a slope nearer to one. This kind of pseudorealism must not be produced at the expense of test score and if the proper emphasis is placed upon score, it probably will not be.

For another example, there is the question of a student's ability to discriminate levels and patterns of uncertainty. To illustrate, consider some data from a 15-item, three-answer test. Figure 18 shows the 15 probability distributions elicited from a student inexperienced in explicitly assessing uncertainty. It appears that this student was thinking in terms of which answer was most likely to be correct and, as a result, responded along the line going from the no-information point up to complete information. Figure 19 shows the 15 probability distributions elicited from a student with considerably more experience in explicitly assessing uncertainty. It appears that this student would sometimes use information to "rule out" one of the answers and perform other kinds of complex discriminations yielding a variety of probability distributions.

Consider now using just one probability distribution to represent each student's knowledge. Let p'_j be the highest probability assigned for item j, p''_j be the next highest, and $p''j'$ the smallest. The average probability distribution ($\bar{p}$, $\bar{p}''$, $\bar{p}'''$) may be found by calculating

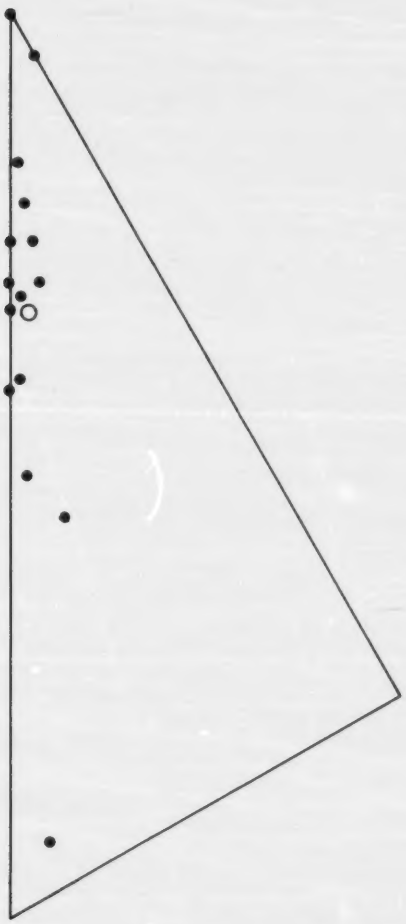


Fig. 18 — Probability distributions (ignoring permutations among the answer labels) used by inexperienced subject taking one 15-item test and yielding a likelihood ratio of .214. Circle represents average probability distribution.

Fig. 19 — Probability distributions (ignoring permutations among the answer labels) used by highly trained subject taking one 15-item test and yielding a likelihood ratio of 36.55. Circle represents average probability distribution.

BEST COPY AVAILABLE

$$\bar{p}' = \frac{1}{n}\sum_{j=1}^{n} p'_j \ , \qquad \bar{p}'' = \frac{1}{n}\sum_{j=1}^{n} p''_j \ , \qquad \text{and} \qquad \bar{p}''' = \frac{1}{n}\sum_{j=1}^{n} p''' \ .$$

This average probability distribution is displayed as a circle in Figs. 18 and 19. Notice that they are not strikingly different for the two students.

Which set of probability distributions, the original set or the average one used for all items, is the better predictor of the set of correct answers? To be more specific, consider the "data" to be the sequence of correct answers and let p_{cj} be the original probability assigned by the student to the correct answer to item j. Then, the likelihood of the data under the hypothesis that they were generated by the student's probability distributions is

$$L_1 = \prod_{j=1}^{n} p''_{cj} \ .$$

Now consider the hypothesis that the data were generated by the constant average probability distribution. That is, look at p_{cj} and give it the value $\bar{p}'$, $\bar{p}''$, or $\bar{p}'''$ according to whether it was the largest, middle, or smallest probability in the set. Or, equivalently, let

n' = the number of times p_{cj} was largest,

n'' = the number of times p_{cj} was next largest, and

n''' = the number of times p_{cj} was the smallest, so that

$$n' + n'' + n''' = n \ .$$

If there are ties among the p_{cj}, fractional numbers must be used. The likelihood of the data under this second hypothesis can be written as

$$L_2 = \bar{p}'^{n'} \bar{p}''^{n''} \bar{p}'''^{n'''} \ .$$

The likelihood ratio can now be computed as L_1/L_2. For the data shown in Fig. 18 this likelihood ratio is about 0.2, indicating that the data are about five times more likely under the constant probability hypothesis. For the data shown in Fig. 19 this likelihood ratio is about 37, indicating that the data were about 37 times more likely using the student's original set of varying probability distributions than using the constant average probability distribution. Thus, this likelihood ratio may prove to be a useful measure of a student's progress in learning how to extract and process information in probabilistic terms.

11. POTENTIAL IMPACT OF TESTING METHOD UPON STUDY BEHAVIOR

Because lower levels of mastery often require much less effort to achieve than do the higher levels, the logarithmic may prove to be a very appropriate reward system that can motivate students to achieve higher levels of mastery of a subject matter than they do at present. To investigate this, assume that the student has, for each question, an exponential "learning curve" of the form

$$p = 1 - \frac{1}{2} \exp(-2\lambda c) ,$$

where c represents the cost to the student in time and energy, say, of the effort he puts into studying the question; λ is a parameter that reflects the "easiness" or rate of learning of the subject matter of the question; and p is the student's probability associated with the correct answer. For the sake of definiteness and simplicity, assume that each question has only two possible answers. Thus, if the student puts no study at all into the question (i.e., $c = 0$), his probability for the correct answer is 0.5, but as he invests effort in studying the subject matter his probability increases asymptotically toward 1.0, as illustrated in Fig. 20.

There are two ways of modeling the way a student will choose to spend his study time and effort. You may either assume that he has a fixed amount of time available and seeks to allocate it across the questions in such a way as to maximize his optimal expected score; or

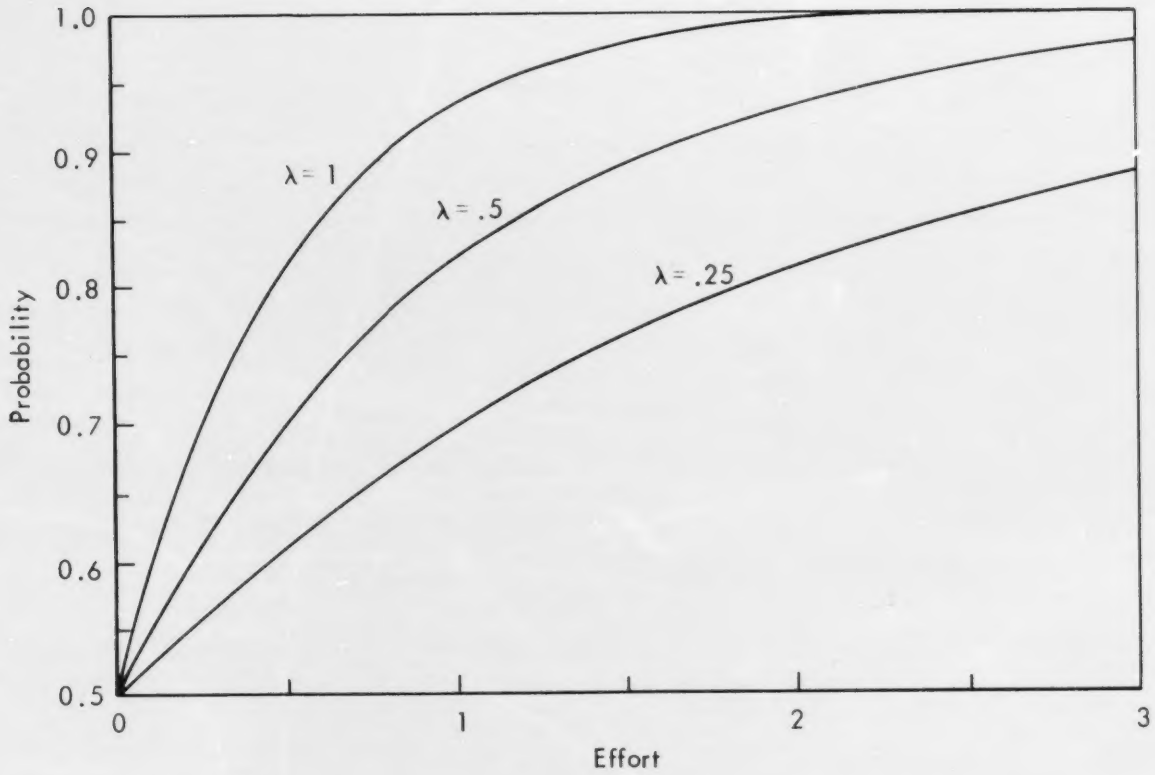


Fig.20 — Probability as a function of effort, c, where $p = 1 - 1/2 \exp(-2\lambda c)$

you may assume that there is some "exchange rate" between study time and score (e.g., one point of score is worth three minutes of time to this particular student) and that he will "spend" his time on each question in such a way as to maximize his "profit," i.e., the difference between his optimal expected score on a question and the value of the time he expends on studying it. These approaches will be discussed separately, but it will become apparent their solutions are closely related.

11.1 Allocation of Study Effort Among Topics

First, suppose that the student has a fixed and limited amount of study time available and wishes to allocate it over the questions likely to be asked in such a way that he will maximize his optimal expected score. On a given question, by following the optimal test-taking strategy he will expect to score

$$E[S(r^*)|p(c)] = E^* ,$$

where p(c) is the function of study time and effort defined in the previous section. Figure 21 shows optimal expected score as a function of effort for a single question under both the simple choice or linear and the logarithmic scoring procedures. The maximum return (in terms of expected score) per unit of effort may be found graphically by measuring the slope of the steepest line through the origin which is tangent to the optimal expected score function, E^*. Analytically, it can be determined by finding the point where the derivative of $(\frac{E}{c})$ with respect to c is zero. Now in fact,

$$\frac{d}{dc}\left(\frac{E}{c}\right) = \frac{1}{c}\frac{dE}{dp}\frac{dp}{dc} - \frac{E}{c^2} = \frac{-(1-p)\log[2(1-p)]\frac{dE}{dp} - E}{c^2} .$$

Because of the particular form chosen for p(c), it follows that the numerator of this expression depends on p alone, not on c or λ. Thus, there exists a "critical value" of p, say p^*, for any given scoring rule such that on any question and regardless of what λ may be, the student will get maximum reward per unit effort to bring his probability for the correct answer up to p^*.

It is easy to calculate p^* for any given scoring rule (see Appendix B). To be specific:

SCORING RULE	CRITICAL PROBABILITY
Simple choice or linear	0.5
Logarithmic	0.891....

An allocation procedure that yields an approximately optimal solution to the overall problem (and an exactly optimal solution in most cases) is as follows. Arrange the questions in order of increasing study difficulty so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The student should work on the first question until he has expended enough effort so that $p \geq p^*$ and the ratio of marginal return to marginal cost (that is, dE/dc) is

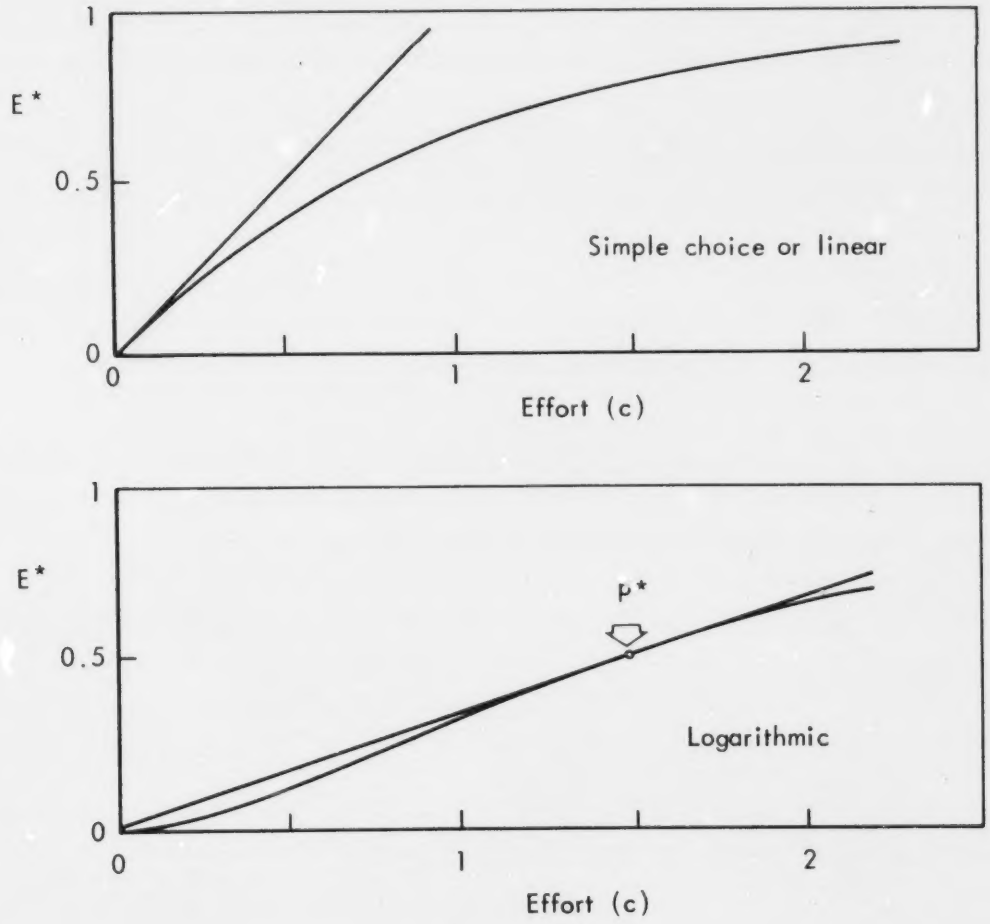


Fig. 21— Optimal expected score as a function of effort (c) when $\lambda = 0.5$

just equal to the maximal achievable gain per unit effort on the second question. Then he should work on the second question until $p \geq p^*$ and then work on the first and second question (keeping marginal return ratios equal) until the marginal return ratios equal the maximal achievable gain per unit effort on the third question. The process is continued until the student has allocated all the effort he has available.

This allocation procedure will yield the true optimum for the scoring rules considered above if the student "runs out of gas" at a point where every question he has worked on at all has been worked on to a point where $p \geq p^*$. In more complex, nonreproducing scoring procedures

that do not have steadily diminishing marginal returns for $p \geq p^*$, the optimal allocation procedure will not work so well.

Now, obviously, a "real-life" student will not go through a careful quantitative analysis of how to allocate his study efforts, but the quantitative model (which may come to represent the behavior of experienced, test-wise students fairly well) does catch one aspect of study behavior that is worth remarking: The use of a logarithmic scoring rule encourages the student to study fewer questions to a higher degree of mastery, while the conventional simple-choice procedure encourages the study of more questions to a lower degree of mastery. Which incentive system is to be preferred depends upon the tradeoffs between scope and retention of the subject matter for the particular learning situation at hand.

Neither incentive system is beyond fault when study time is strictly limited. On the one hand, use of the conventional simple-choice procedure may mean that the student will remember none of the subject matter more than a few hours or days after he takes the test. On the other hand, if he uses the logarithmic procedure he may remember some of the subject matter, but not enough for it to be of any use to him. "Cramming" for a test can easily be a losing proposition which, with the simple-choice procedure, yields an adequate test score but produces little learning.

11.2 Investment of Study Effort in a Single Topic

An alternative way of modeling the student's study incentives is to assume that his study time is not strictly limited and that his time has a value to him which is commensurable to the value of the test score he may earn. If the total amount of time which he may spend on study is flexible, he would perhaps attempt to maximize his "profit" on each test question. That is to say, he would choose an expenditure of time c^* on each question that maximizes $E[r^*|p(c)] - sc$, where s is the value, in units of test score, of a single unit of time (or study effort). Assume for the moment that the units of time (or study effort) have been normalized in such a way that $s = 1$.

Within the context of the quantitative model it is an easy task to calculate (see Appendix C) as a function of λ, the optimal investment

strategy and maximal point under both the simple choice and the logarithmic scoring rules. The results of these calculations are graphed in Fig. 22. For a given λ the simple choice procedure allows the larger profit and, in this sense, is a more lenient reward system than is the logarithmic. Under the simple choice procedure it never pays to work on a question where $\lambda < 0.5$, while under the logarithmic the student cannot make a profit if $\lambda < 1.5$. If $\lambda \geq 1.5$, the student will expend considerably more effort under the logarithmic scoring rule. Note, by the way, that if the student studies a question at all under the "maximum profit" hypothesis, he studies it at least up to the level where his probability exceeds p^*.

Thus, the same basic pattern appears under the "maximum profit" hypothesis as under the "optimal allocation" hypothesis. Specifically, the student is theoretically motivated to study fewer questions (through avoidance of the harder ones with $\lambda < 1.5$) but to a higher degree of mastery under the logarithmic scoring rule than under the conventional simple choice procedure. In the case of the investment problem, however, the student may be induced to study all of the questions by increasing the reward for learning or by increasing the rate of learning (λ) either through improving learning efficiency or through reorganization of the subject matter. Any of these steps may serve to resolve the conflict between scope of learning and retention.

Whether these effects will be observable in real students in real-life situations will be an interesting matter to investigate empirically.

12. IMPACT OF INAPPROPRIATE REWARDS UPON TEST-TAKING BEHAVIOR

A fundamental assumption underlying all of the above analyses of optimal behavior is that the student wishes to maximize his expected test score. What may happen when this condition is relaxed?

With the simple choice procedure, a student desiring to maximize expected test score does it by selecting, for each question on the test, that answer he considers most likely to be correct, as shown in Sec. 5.1. Suppose, however, that a cutting score or some grading limits are imposed on the test so that the student now wishes to maximize the probability that his test score will equal or exceed a specified score, say N or more answers correct.

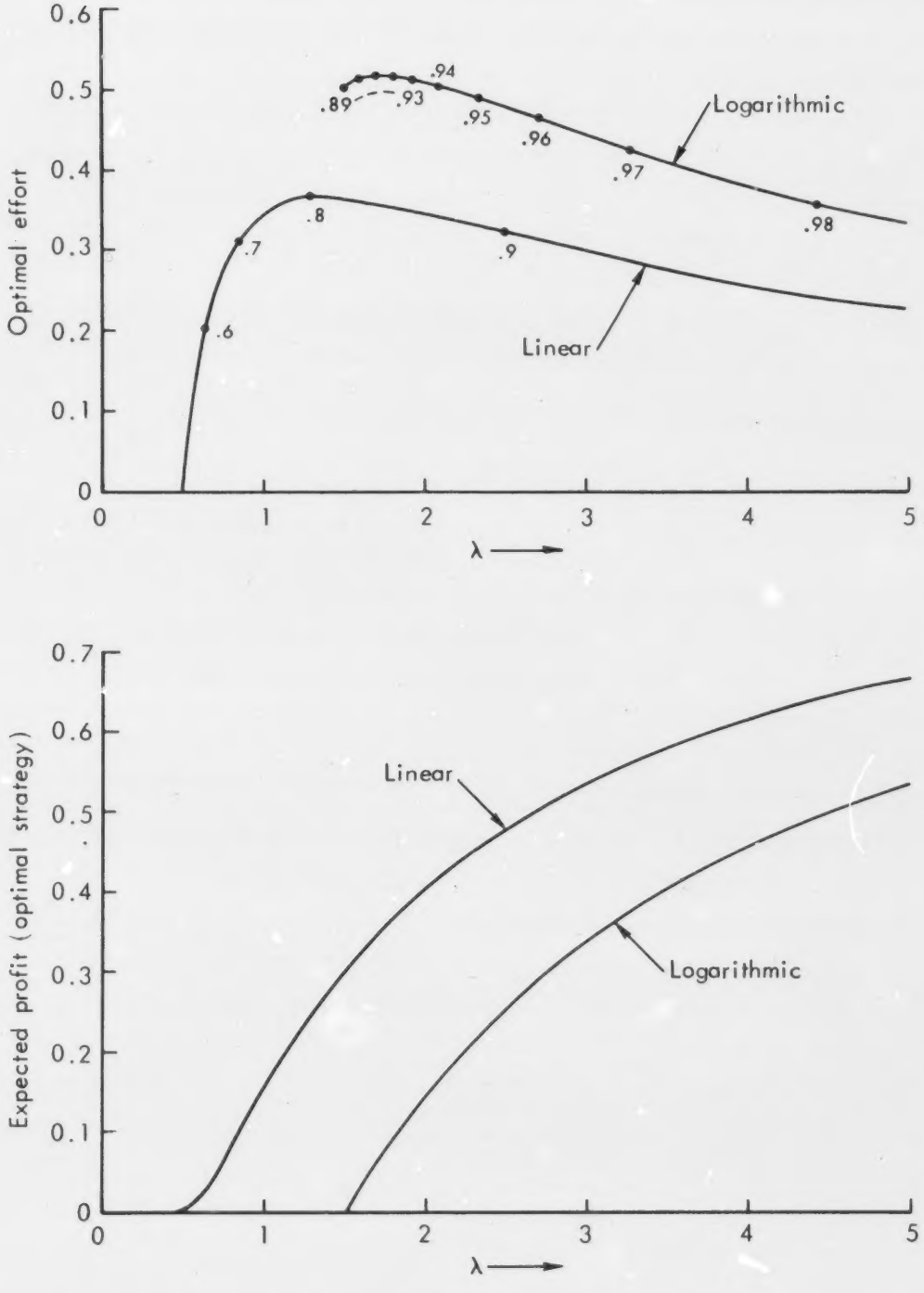


Fig. 22— Optimal investments and profits as a function of rate of learning

To find the optimal test-taking strategy under this reward structure, assume that the student perceives all the questions to be independent. That is to say, he feels that the probable correctness of the answers to one question would not be affected by what the correct answer turns out to be on another question. Now let

$p_{i(j),j}$ = probability of getting question j correct given that student chooses answer i(j),

$p_{\ell}(K)$ = probability of getting K correct out of the first ℓ questions,

$p_{\ell}(K+)$ = probability of getting K or more correct out of the first ℓ questions.

Then,

$$p_n(N+) = \sum_{h=N}^{n} P_n(h)$$

$$= \sum_{h=N}^{n} \left\{ p_{i(n),n}\, P_{n-1}(h-1) + [1 - p_{i(n),n}]\, P_{n-1}(h) \right\}$$

$$= p_{i(n),n}\, P_{n-1}(N-1) + P_{n-1}(N+) \ .$$

Since $P_{n-1}(N-1) \geq 0$, regardless of what strategy the student uses on the first n - 1 questions, it follows that choosing i(n) so that $p_{i(n),n}$ will be a maximum will give the student an equal or better chance of getting N or more correct as will any other choice on the n*th* question. Clearly, the questions could be renumbered to make any question the "n*th* question," and thus the obvious strategy is, indeed, an optimal one.

The assumption of independence among the test items was used in the proof given above. Consider now an example showing that this result does, in fact, depend on the assumption of independence. Here is the test:

1. It rained in Santa Monica on July 24, 1932. True or False?
2. It did not rain in Santa Monica on July 24, 1932. True or False?

You must get at least one item right to pass the test. Obviously, if you answer both items "True" or both items "False" you are certain to pass. If you are 90 percent certain that it did *not* rain in Santa Monica on July 24, 1932 and you use the "obvious" strategy, then there is a 10 percent chance that you will flunk. This shows that the obvious strategy is not necessarily optimal if the questions are not independent.

Be that as it may, the simple-choice procedure is relatively insensitive to the reward structure within which it is embedded. As a consequence of this property of the widely used simple-choice scoring procedure, test givers have probably gotten in the habit of ignoring reward structures and can afford to use cutoff scores and prizes with abandon. Such behavior can cause great difficulty when one attempts to improve testing through the elicitation of personal probabilities.

The notion that the student should answer each question in such a way as to maximize his expected score is based upon the assumption that he has a linear utility for points. In many educational contexts as they currently exist, this assumption will be manifestly out of line with the facts.

For example, suppose that some special prize is to be given to whoever gets the best score for a given test. This will tend to make students overstate their probabilities (or, to put it another way, to appear to overvalue their information), because the chance of getting a really high score will be worth more than the risk of getting an unusually low score (which will be no worse for the student than a mediocre score). The precise quantitative measurement of this effect is very difficult in general, because it involves a multiperson game that is affected not only by each player's perception of the difficulty of the questions but also by his perception of the ability of the other players. However, an analysis of what happens if two players are asked a single question will be found in [7], pp. 12-13.

The special case in which a prize is awarded only in the event that the student makes a perfect score is very easy to understand. With this

BEST COPY AVAILABLE

reward structure, the student should always set one of the $r_i = 1$ no matter how great his uncertainty because if he fails to do so, he will foreclose any possibility of making a perfect score.

Another context in which students might be motivated to give responses other than their personal probabilities is any situation in which all that matters is to achieve a given level of score. For example, if the students are on a "pass-fail" system, where they pass the course if they achieve a certain test score or better, and fail the course otherwise, then they may have considerable incentive to shade their responses up or down from their probabilities. The general problem of determining an optimal response strategy under these circumstances is mathematically very complex and no solution is known. The following simplified example, however, can be solved and it illustrates very clearly how the imposition of a "pass-fail" reward structure on top of a reproducing scoring system may completely destroy any incentive for students to respond with their probabilities.

Suppose a student faces an exam consisting of n two-answer items. Suppose these questions all "look alike" to the student, in the sense that on each question he has a fixed probability distribution, p and $1 - p$, with $p \geq 1/2$. Suppose that he requires a total score T on the test in order to pass. He wants to choose a fixed response r to assign to the preferred answer to each question. What value of r should he choose in order to maximize his probability of passing the test? It is not hard to show (see Appendix D), that he will have the maximal probability of passing if he chooses r such that $E[S(r)|r] = T/n$. Note that this r does not depend on p at all! So the student's optimal test-taking strategy depends only on what score he must make in order to pass, and not on his level of knowledge with respect to each test item. In short, this reward structure utterly destroys the reproducing character of the scoring rule. Figure 23 illustrates the student's probability of passing as a function of his response strategy in the particular case where $T = 0.58$, $n = 20$, and $p = 0.8$. Note that the student will be about nine times as likely to fail the test if he pursues the "maximum expected value" strategy as he will be if he follows the "maximum probability of passing" strategy.

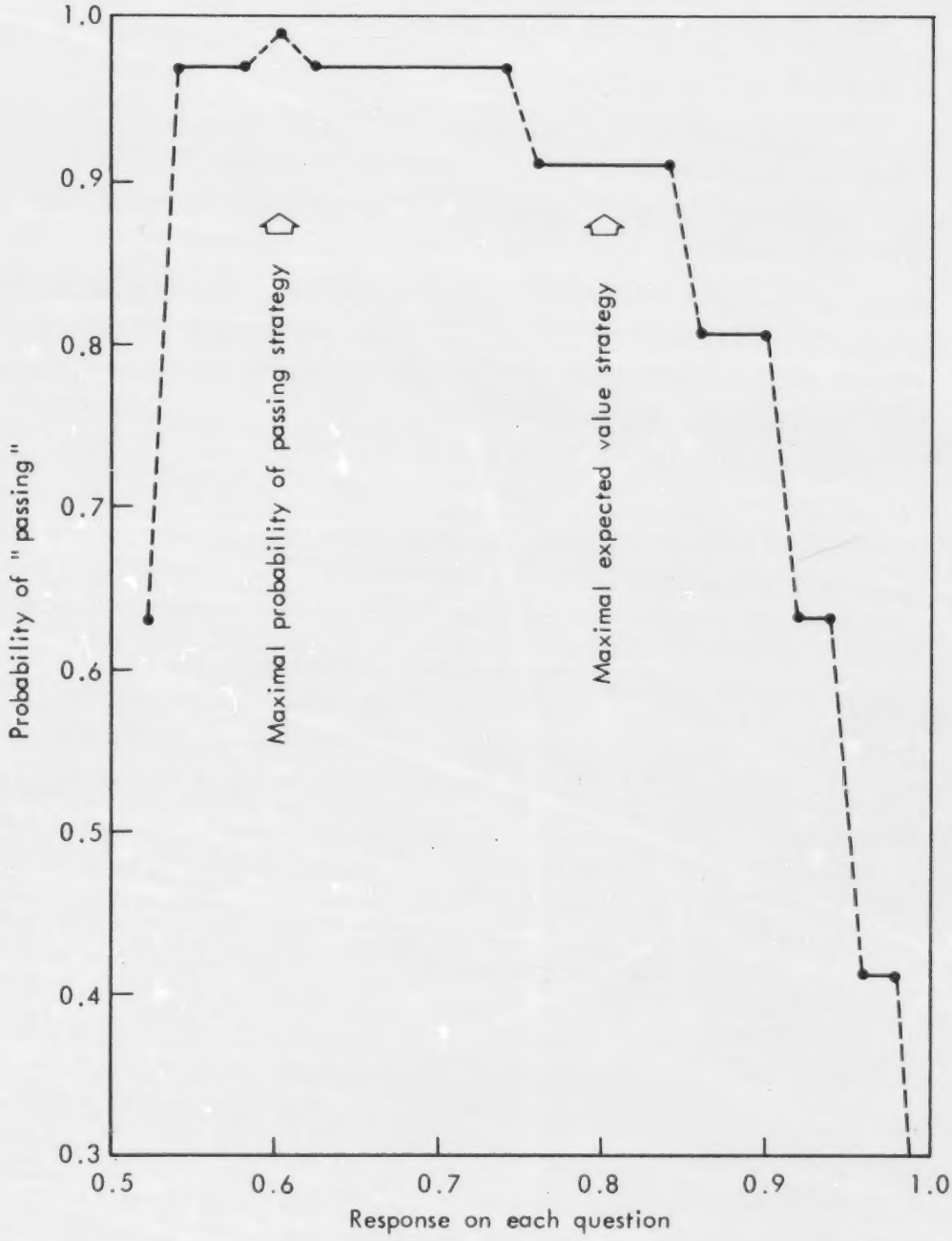


Fig. 23 — Probability of "passing" as a function of response to each question in a 20 question test; Logarithmic scoring system; probability 0.8 on each question; Score required to pass T = 0.58

In an actual situation, however, the reproducing character of the scoring rule would not be completely washed out, because the student would *not* have precisely the same probability distribution for each item. It seems intuitively evident (although a rigorous proof has not yet been discovered) that his best strategy would be to hedge all his responses but still let his responses vary somewhat with his probabilities.

But the best remedy is to avoid creating reward structures which put a highly nonlinear value on points earned under an allegedly reproducing scoring rule. Another (partial) remedy is to avoid letting the student know how many questions there are on a test, or how difficult they are, before he begins to take it.

13. SUMMARY AND CONCLUSIONS

We have seen that it is patently desirable to broaden the responses that students are permitted to make to multiple-choice questions. The reasons for this are as follows: the student is then able to transmit more information to the teacher on each item; conventional multiple-choice tests do nothing to train the student to weight the strength of conviction justified by his knowledge on a given item; and students themselves prefer greater freedom of response and chafe under the limitations of the conventional one-choice response format.

However, it is meaningless or even deceptive to permit students to give a weighted response rather than a unitary choice if the scoring system is not carefully chosen so as to encourage students to use the full range of choice available to them. For example, if the student is allowed to respond with weights (which add up to one over all alternative responses on each question) and is then given a score on each question equal to the weight he ascribed to the correct alternative, then it will not take an intelligent student long to recognize that he should not utilize the freedom you have made available to him, but simply respond with weights of zero and one as in a conventional multiple-choice test. One excellent solution to this problem appears to be the use of "admissible scoring systems," which are designed to provide the student with a maximum expected score if he makes his responses correspond to his subjective probabilities.

Admissible scoring procedures have many desirable features. They link the student's responses to the well-developed disciplines of subjective probability, information theory, and Bayesian decisionmaking. The student who becomes "test-wise" against a reproducing scoring system has learned to express his uncertainty in the universal language of probability theory. He has also learned to weight the facts, clues, and reasons available to him and come up with a "risk-balancing" response. Preliminary data from computer-administered admissible probability testing show that, while some people possess this aptitude, others are quite biased in their assessment of uncertainty and could benefit greatly from further training in this skill. Admissible scoring procedures also have the theoretical advantage that they lead the student toward higher degrees of mastery than do conventional scoring procedures. That is to say, the student perceives increased rewards for higher degrees of certainty on each question under an admissible scoring system than under a conventional multiple-choice scoring system. The latter tends to encourage superficial knowledge of a wide variety of topics; the former encourages total mastery of a smaller number of topics. This perception should have a desirable effect on the student's study habits. Whether this effect will be observed in practice makes an interesting topic for future experiments.

It will be very important, in practical applications of admissible probability testing, to insure that the *external* incentive system (i.e., what is done with the test scores) be consistent with the basic assumption of admissible probability testing. That is to say, the students must perceive the maximization of expected score as being their best strategy. In theory, the use of a "pass-fail" system, or the use of extreme competition, may have the effect of distorting the students' responses away from their true subjective probabilities. The value of the students' rewards must be somehow proportional to the total score they each receive. Whether this problem turns out to be serious or not is another question that can be answered only by empirical tests.

Appendix A

FITTING A PLANAR REALISM FUNCTION

1. NOTATION, NORMALIZATION, AND SYMMETRY

Assume there are n questions in the test, with three possible answers for each question. We reorder the answers so that $p_1^j \geq p_2^j \geq p_3^j$, where p_i^j is the probability the student ascribes to the i*th* answer on the j*th* question. We let i_j denote the *correct* answer on the j*th* question.

We wish to find a linear transformation

$$\begin{aligned} q_1 &= a_{11}p_1 + a_{12}p_2 + a_{13}p_3 \\ q_2 &= a_{21}p_1 + a_{22}p_2 + a_{23}p_3 \\ q_3 &= a_{31}p_1 + a_{32}p_2 + a_{33}p_3 \end{aligned} \tag{A.1}$$

that will minimize the quantity,

$$\sum_{j=1}^{n} \sum_{i=1}^{3} (q_i^j - e_i^j)^2 = \Delta \ , \tag{A.2}$$

where e_i^j is zero or one depending upon whether answer i to the j*th* question is incorrect or correct.

In addition to minimizing expression (A.2), we require the transformation to meet certain conditions of normality and symmetry:

(A) If $\sum_{i=1}^{3} p_i = 1$, then $\sum_{i=1}^{3} q_i = 1$.

(B) If $p_1 = p_2$, then $q_1 = q_2$.

(C) If $p_2 = p_3$, then $q_2 = q_3$.

All three of these conditions together imply that $p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is carried into $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ by our transformation; thus for all i,

$$\sum_{k=1}^{3} a_{ik} = 1 . \tag{A.3}$$

Condition A, applied in turn to $p = (1,0,0)$, $p = (0,1,0)$ and $p = (0,0,1)$, implies that for all i,

$$\sum_{k=1}^{3} a_{ki} = 1 . \tag{A.4}$$

Condition B, applied to $(\frac{1}{2}, \frac{1}{2}, 0)$, implies

$$a_{11} + a_{12} = a_{21} + a_{22} . \tag{A.5}$$

Condition C, applied to (1,0,0), implies

$$a_{21} = a_{31} . \tag{A.6}$$

Now let us denote a_{11} by α, and a_{13} by β. From (A.6) and (A.4) we see that

$$a_{21} = a_{31} = \frac{1 - \alpha}{2} . \tag{A.7}$$

From (A.3) we see that

$$a_{12} = 1 - \alpha - \beta . \tag{A.8}$$

From (A.5), (A.7), and (A.8) we have

$$a_{22} = \frac{1 + \alpha - 2\beta}{2} . \tag{A.9}$$

From (A.8), (A.9), and (A.3) we derive

$$a_{23} = \beta \ . \tag{A.10}$$

Applying (A.4) now yields

$$a_{32} = \frac{-1 + \alpha + 4\beta}{2} \quad \text{and} \quad a_{33} = 1 - 2\beta \ . \tag{A.11}$$

In summary, the application of normality and symmetry conditions shows that system (A.1) may be written

$$\begin{aligned} q_1 &= \alpha p_1 + (1 - \alpha - \beta)\, p_2 + \beta p_3 \\ q_2 &= \frac{1 - \alpha}{2}\, p_1 + \frac{(1 + \alpha - 2\beta)}{2}\, p_2 + \beta p_3 \\ q_3 &= \frac{1 - \alpha}{2}\, p_1 + \frac{(-1 + \alpha + 4\beta)}{2}\, p_2 + (1 - 2\beta) p_3 \ . \end{aligned} \tag{A.12}$$

It is easy to see that these expressions are necessary and sufficient conditions for (A), (B), and (C) to hold.

The parameters α and β have an immediate interpretation, as follows. The requirement that $p_1 \geq p_2 \geq p_3$ means that we are restricting our attention to one-sixth of the "answer triangle" (the shaded area in the upper left-hand triangle of Fig. 24). The mapping (A.12) leaves the point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ fixed, carries $(1,0,0)$ into $(\alpha,0,0)$, and carries $(\frac{1}{2}, \frac{1}{2},0)$ into $(\frac{1 - \beta}{2}, \frac{1 - \beta}{2}, \beta)$. These three points are the vertices of the shaded triangle, and by knowing what happens to them it is easy to visualize what happens to all other points in the triangle. Figure 24 includes three examples of what the mappings look like for different values of α and β.

2. MINIMIZATION OF Δ

This section gives the formulas required to calculate values of α and β that will minimize Δ, the quantity defined by (A.2). The derivation of these formulas is by taking the derivative of Δ with respect

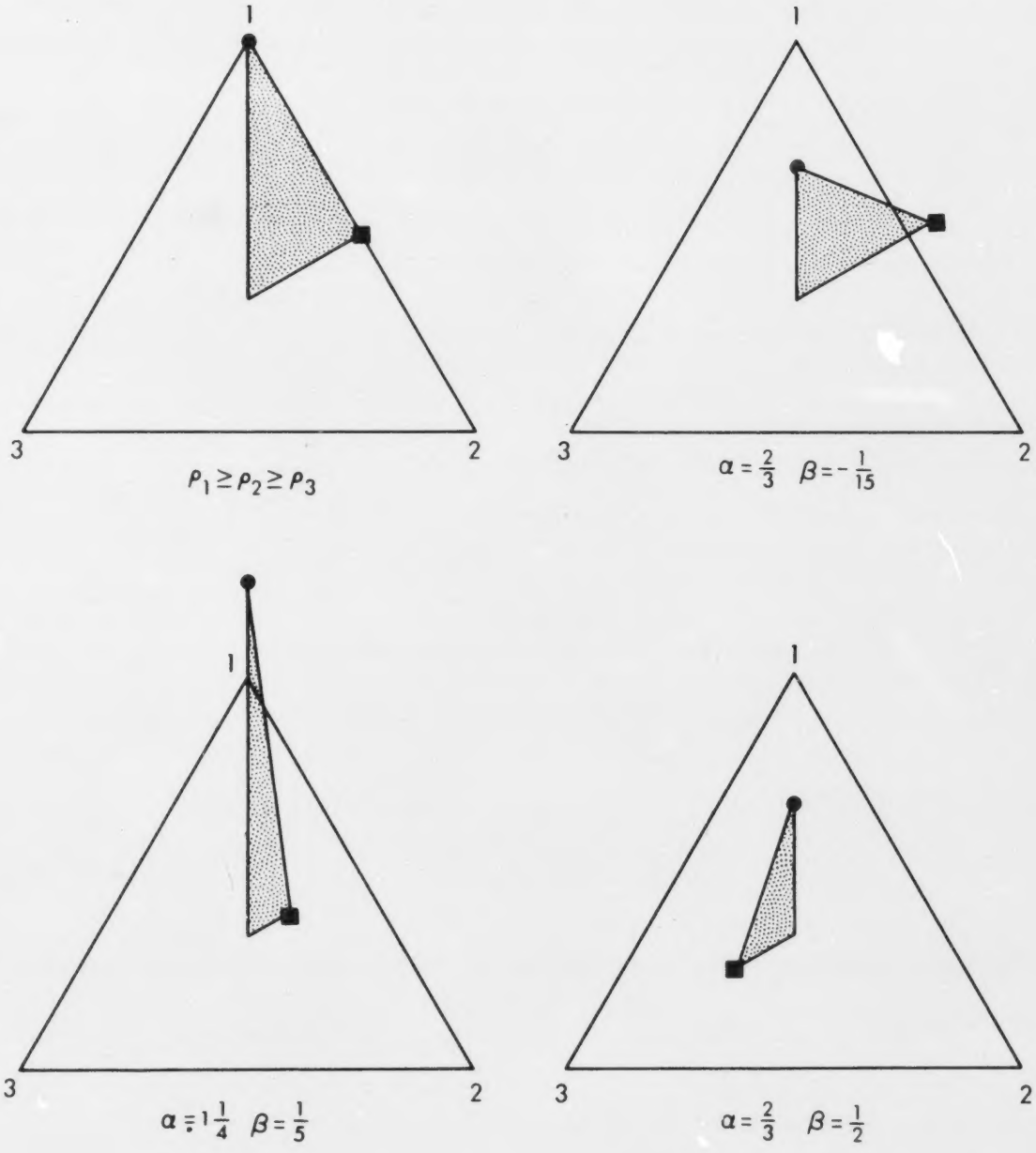


Fig. 24— Effect of the linear transformation on the triangle $\rho_1 \geq \rho_2 \geq \rho_3$ for various values of α and β

to α and β. We skip the intermediate steps in this routine calculation, and jump directly to our final formulas:

$$\alpha = \frac{c_{\alpha.}\, c_{\beta\beta} - c_{\beta.}\, c_{\alpha\beta}}{c_{\alpha\alpha}\, c_{\beta\beta} - c_{\beta\alpha}\, c_{\alpha\beta}} ;$$

$$\beta = \frac{c_{\beta.}\, c_{\alpha\alpha} - c_{\alpha.}\, c_{\beta\alpha}}{c_{\alpha\alpha}\, c_{\beta\beta} - c_{\beta\alpha}\, c_{\alpha\beta}} . \qquad \text{(A.13)}$$

The quantities appearing in these formulas are defined as follows. Let

$$A_1^j = p_1^j - p_2^j \qquad B_1^j = p_3^j - p_2^j \qquad c_1^j = p_2^j$$

$$A_2^j = \frac{p_2^j - p_1^j}{2} \qquad B_2^j = p_3^j - p_2^j \qquad C_2^j = \frac{p_1^j + p_2^j}{2} \qquad \text{(A.14)}$$

$$A_3^j = \frac{p_2^j - p_1^j}{2} \qquad B_3^j = 2(p_2^j - p_3^j) \qquad C_3^j = \frac{p_1^j - p_2^j}{2} + p_3^j$$

The reader will note that

$$q_i^j = A_i^j\, \alpha + B_i^j\, \beta + C_i^j \qquad \text{(A.15)}$$

Now we define

$$c_{\alpha\alpha} = \sum_{j=1}^{n} \sum_{i=1}^{3} A_i^j\, A_i^j \qquad c_{\alpha\beta} = c_{\beta\alpha} = \sum_{j=1}^{n} \sum_{i=1}^{3} A_i^j\, B_i^j$$

$$c_{\beta\beta} = \sum_{j=1}^{n} \sum_{i=1}^{3} B_i^j\, B_i^j$$

$$c_{\alpha.} = \sum_{j=1}^{n} A_{i\,j}^{\,j} - \sum_{j=1}^{n} \sum_{i=1}^{3} A_i^j\, C_i^j \qquad \text{(A.16)}$$

$$c_{\beta.} = \sum_{j=1}^{n} B_{i\,j}^{\,j} - \sum_{j=1}^{n} \sum_{i=1}^{3} B_i^j\, C_i^j$$

Recall that i_j is the subscript of the probability ascribed to the correct answer on the j*th* question. Now the reader may be bothered by the following question: suppose a respondent on the first questions lists probabilities (0.4, 0.2, 0.4), and the third answer is in fact correct. We reorder the probabilities to get (0.4, 0.4, 0.2), but what value do we take for i_1? Should it be 1 or 2? It does not matter, as far as calculating our coefficients α and β is concerned, for i_j enters into the calculation only as a subscript for A's and B's; when this ambiguity arises about whether $i_j = 1$ or $i_j = 2$, we have $p_1^j = p_2^j$ and so $A_1^j = A_2^j$ and $B_1^j = B_2^j$. Similarly, if there is ambiguity over whether $i_j = 2$ or $i_j = 3$, we have $A_2^j = A_3^j$ and $B_2^j = B_3^j$. Because of the symmetry built into Eq. (A.12), it also makes no difference which possible value of i_j we select (where ambiguity exists) in calculating the total score awarded to the "transformed" estimates.

3. TRUNCATION AND RENORMALIZATION

The procedure above does not necessarily lead to a vector (q_1, q_2, q_3) that is a proper probability vector. Although the q's will sum to one, they will not necessarily fall between zero and one. We truncate and renormalize in the obvious way:

$$q_i^{j*} = \min \frac{(1, \max (0, q_i^j))}{d_j} \qquad (A.17)$$

$$d_j = \sum_{i=1}^{3} \min (1, \max (0, q_i^j)) .$$

This truncation may seem rather arbitrary and ad hoc. Recall, however, that it will take place only if the respondent *under*estimates his knowledge (and α and β do not fall between zero and one), a phenomenon that so far has occurred only rarely and with naive subjects. Therefore, the use of a more sophisticated truncation and renormalization routine hardly seems justified.

BEST COPY AVAILABLE

Appendix B

HOW TO CALCULATE THE VALUE OF "p" AT WHICH MAXIMUM EXPECTED RETURN PER UNIT EFFORT IS ACHIEVED

Recall that in Sec. 11 we assumed, for a given true-false question, that a student's probability, p, of choosing the correct response could be expressed as an exponential function of the effort he put into studying the question. Specifically, we assumed that

$$p = 1 - \frac{1}{2}\exp(-2\lambda c) \ , \tag{B.1}$$

where c represents the study-time ("cost") and λ is a parameter reflecting the "easiness" of the question. Recall also that the expected score a student is able to make on a question can be expressed as a function of his probability of choosing the correct response. Specifically,

$$E[S(p)\,|p] = pS(p) + (1 - p)S(1 - p) \ . \tag{B.2}$$

We assume in the formula (B.2) that the scoring function is symmetric; i.e., the student gets exactly as much credit for "0.7 true; 0.3 false" if "true" is correct as he gets for "0.3 true; 0.7 false" if "false" is correct. By combining (B.1) with (B.2) we may express maximum expected score directly as a function of "cost," c, and "easiness," λ:

$$E[\lambda,c] = (1 - \tfrac{1}{2}\exp(-2\lambda c))S(1 - \tfrac{1}{2}\exp(-2\lambda c))$$

$$+ \tfrac{1}{2}\exp(-2\lambda c)S(\tfrac{1}{2}\exp(-2\lambda c)) \ . \tag{B.3}$$

Now, it is immediately evident that $E[1,\lambda c] = E[\lambda,c]$ for all positive values of λ and c. If we are looking for the maximum return per unit effort, we may apply this observation as follows:

$$\max_{c\geq 0} \frac{E(\lambda,c)}{c} = \max_{c\geq 0} \lambda \frac{E(1,\lambda c)}{\lambda c} = \lambda \max_{c\geq 0} \frac{E(1,c)}{c} \qquad (B.4)$$

In other words, if c_λ^* represents the cost that maximizes $\frac{E(\lambda,c)}{c}$, then $\lambda c_\lambda^* = c_1^*$. This is true, in an obvious sense, even if the c_λ^* are not unique. In our learning model, p depends on the product of λ and c. This is extremely convenient, for if we let $p_\lambda^* = \lambda c_\lambda^*$, we see that $p_\lambda^* = p_1^*$. In other words, the maximum return per unit effort is achieved on a given true-false question by studying that question until a given probability of choosing the correct answer is achieved; and this "mastery level" (which we shall call p^*) does not depend on the easiness of the question, but only on the scoring function used. This critical mastery level is thus a characteristic of the scoring function; presumably students will study harder when faced with a scoring function with a high critical mastery level than they will when faced with a scoring function having a low one.

If the scoring function is differentiable, elementary calculus may be used to calculate the value of p^*. One good way to do this is to use the chain rule, as follows:

$$\frac{d}{dc}\left(\frac{E[S(p)|p]}{c}\right) = \frac{1}{c}\frac{dE}{dp}\frac{dp}{dc} = \frac{E}{c^2} . \qquad (B.5)$$

Example 1: Let $S(p) = \frac{\log(2p)}{\log(2)}$. This is the logarithmic scoring rule normalized so that $S(\frac{1}{2}) = 0$ and $S(1) = 1$. Then

$$E = \frac{p \log(2p) + (1 - p) \log(2(1 - p))}{\log 2}$$

$$\frac{dE}{dp} = \frac{\log(\frac{p}{1 - p})}{\log 2} \qquad (B.6)$$

$$\frac{dp}{dc} = \lambda \exp(-2\lambda c) = 2(1 - p)\lambda .$$

Therefore,

$$c^2 \log(2) \frac{d}{dc}\left(\frac{E}{c}\right) = c \cdot \left[\log\left(\frac{p}{1-p}\right)\right] \cdot 2(1-p)\lambda$$

$$-p \log(2p) + (1 - p \log(2(1-p)) \ . \qquad \text{(B.7)}$$

Since

$$c = \frac{-\log[2(1-p)]}{2} \ , \qquad \text{(B.8)}$$

we see that (B.7) may be expressed as

$$c^2 \log 2 \frac{d}{dc}\left(\frac{E}{c}\right) = -\log\left(\frac{p}{1-p}\right) \log(2(1-p))(1-p)$$

$$- p \log 2p - (1-p) \log(2(1-p)) \ . \qquad \text{(B.9)}$$

The maximum value of $\frac{E}{c}$ will be achieved at the point where the right-hand side of (B.9) equals zero. Solving this transcendental equation is very difficult by hand, but is easy (using the method of false position or Newton's method) on any computer. The derivative in (B.9) is zero at $p = \frac{1}{2}$; is positive for $\frac{1}{2} < p < 0.8910751 \ldots$; and is negative for $0.8910751 \ldots < p$. Thus the maximum expected score per unit effort is achieved for $p = 0.8910751 \ldots$.

Example 2: Let $S(p) = 1 - 4(1-p)^2$. This is the "quadratic scoring system," or "Brier score," often used by meteorologists to evaluate the quality of probabilistic weather predictions. In our case, we normalize it so that $s(\frac{1}{2}) = 0$ and $s(1) = 1$. Then

$$E = (2p-1)^2$$

$$\frac{dE}{dp} = 8p - 4 \qquad \text{(B.10)}$$

$$\frac{dp}{dc} = 2(1-p)\lambda \ .$$

So

$$c^2 \frac{d}{dc}\left(\frac{E}{c}\right) = c \cdot (8p - 4)\, 2 \cdot (1 - p)\lambda - (2p - 1)^2 . \qquad (B.11)$$

Using relation (B.8), we see that

$$c^2 \frac{d}{dc}\left(\frac{E}{c}\right) = \left[-8 \cdot \log[2 \cdot (1 - p)] \cdot (1 - p) - (2p - 1)\right](2p - 1) . \qquad (B.12)$$

Simple calculation shows that the derivative in (B.12) is zero at $p = \frac{1}{2}$; positive for $\frac{1}{2} < p < 0.857665933 \ldots$; and negative for $p > 0.85766593 \ldots$. Thus the maximum expected score per unit effort against the quadratic scoring system is achieved at $p = 0.85766593 \ldots$. In short, the quadratic scoring system is apparently slightly less effective (theoretically) than the logarithmic scoring system in stimulating students to work hard on individual questions.

Example 3: The techniques of this appendix may also be applied to scoring systems that are not admissible. For example, suppose a student is approaching a true-false test that is to be marked and graded in the traditional way (+1 for a right answer; -1 for a wrong one). Then if a student has probability p of selecting the right answer, his expected score on the question will be $p \cdot (+1) + (1 - p) \cdot (-1)$. We therefore have

$$E = 2p - 1$$

$$\frac{dE}{dp} = 2 \qquad (B.13)$$

$$\frac{dp}{dc} = 2(1 - p)\lambda .$$

Thus

$$c^2 \frac{d}{dc}\left(\frac{E}{c}\right) = c \cdot 2 \cdot 2(1 - p)\lambda - (2p - 1) . \qquad (B.14)$$

It follows that

$$c^2 \frac{d}{dc}\left(\frac{E}{c}\right) = -2(1 - p)\log(2(1 - p)) - (2p - 1) . \qquad (B.15)$$

The derivative in (B.15) is negative for $\frac{1}{2} < p < 1$. It follows that $\left(\frac{E}{c}\right)$ is a maximum at $p = \frac{1}{2}$. In other words, the maximum return per unit effort against a conventional true-false question is achieved by putting forth an infinitesimal amount of effort.

Appendix C

ALLOCATING STUDY EFFORT TO MAXIMIZE PROFIT

In Appendix B we analyzed the problem of how much study would get the maximum return per unit effort. Another approach to the question of how students will be motivated to study is to suppose that study effort and points gained on a test question can be measured in commensurable terms. The student is then in the position of "purchasing" expected score with study effort. You might expect the student to attempt to maximize his "profit"; that is, to try to make the difference between the value of the score he expects to gain and the value (to him) of the effort he expends in study. In short, he will try to maximize

$$E(\lambda,c) - c \ . \tag{C.1}$$

If the reward system is such that E is a differentiable function of p, then this maximization problem may be solved by finding the point at which the derivative is zero.

$$\frac{d}{dc}(E - c) = \frac{dE}{dp}\frac{dp}{dc} - 1 \ . \tag{C.2}$$

We assume, as in Appendix B, that

$$\begin{aligned} p &= 1 - \frac{1}{2}\exp(-2\lambda c) \\ \frac{dp}{dc} &= \lambda \exp(-2\lambda c) = \lambda 2(1 - p) \ . \end{aligned} \tag{C.3}$$

Combining (C.2) with (C.3) we see that the derivative of profit will be zero where

$$\lambda = \frac{1}{2(1 - p)\frac{dE}{dp}} \ . \tag{C.4}$$

Although we would ordinarily think of fixing λ and then solving for p (or what is the same thing, c), Expression (C.4) is such an easy formula that the best way to derive numerical values seems to be to regard p as a parameter and derive the maximum expected profit as a function of λ by plotting the curve $\left(E(p) + \frac{\log 2(1-p)}{2\lambda(p)}, \lambda(p)\right)$.

Example 1: Consider the logarithmic scoring system, $S(p) = \frac{\log 2p}{\log 2}$. Then we have

$$\lambda(p) = \frac{\log 2}{2(1-p)[\log(p/1-p)]}$$

$$E(p) - c = 1 + \frac{p \log p + (1-p)\log(1-p)}{\log 2} \qquad (C.5)$$

$$+ \frac{(1-p)\log(p/1-p)\log 2(1-p)}{\log 2} .$$

Carrying out these calculations yields the following values:

p	lambda	cost	profit
0.99	7.542	0.25934	0.65986
0.98	4.453	0.36146	0.49710
0.97	3.323	0.42327	0.38233
0.96	2.726	0.46321	0.29449
0.95	2.354	0.48906	0.22454
0.94	2.099	0.50500	0.16756
0.93	1.914	0.51360	0.12048
0.92	1.774	0.51658	0.08124
0.91	1.664	0.51514	0.04839
0.90	1.577	0.51018	0.02082
0.89	1.507	0.50238	-0.00229
0.88	1.450	0.49226	-0.02163

Note that for p less than about 0.9 there is really no value of λ for which such a p is optimal, since it is better not to study at all (and get zero profit) than to do any studying and get a negative profit.

Example 2: Now let us turn to the quadratic scoring system, $S(p) = 1 - 4(p-1)^2$.

$$\lambda(p) = \frac{1}{2(1 - p)[4(2p - 1)]}$$

$$E(p) - c = (1 - 2p)^2 + (1 - p)4(2p - 1)\log 2(1 - p) . \qquad (C.6)$$

Carrying out these calculations leads to the following:

p	lambda	cost	profit
0.99	12.755	0.15335	0.80705
0.98	6.510	0.24721	0.67439
0.97	4.433	0.31735	0.56625
0.96	3.397	0.37179	0.47461
0.95	2.778	0.41447	0.39553
0.94	2.367	0.44780	0.32660
0.93	2.076	0.47344	0.26616
0.92	1.860	0.49260	0.21300
0.91	1.694	0.50621	0.16619
0.90	1.563	0.51502	0.12498
0.89	1.457	0.51965	0.08875
0.88	1.371	0.52061	0.05699
0.87	1.299	0.51835	0.02925
0.86	1.240	0.51326	0.00514
0.85	1.190	0.50567	-0.01567

If λ is less than about 1.2, it is better not to study at all, and accept zero profit, for no finite amount of effort expended will lead to a commensurate reward.

Example 3: As a final example, consider a normal true-false test. This is not an admissible scoring system, but we can see that $E(p) = 2p - 1$ $(p \geq \frac{1}{2})$. Thus

$$\lambda(p) = \frac{1}{4(1 - p)}$$

$$E(p) - c = 2p - 1 + 2(1 - p)\log 2(1 - p) . \qquad (C.7)$$

Computation yields the following:

p	lambda	cost	profit
0.95	5.000	0.23026	0.66974
0.90	2.500	0.32189	0.47811
0.85	1.667	0.36119	0.33881
0.80	1.250	0.36652	0.23348
0.75	1.000	0.34657	0.15343
0.70	0.833	0.30650	0.09350
0.65	0.714	0.24967	0.05033
0.60	0.625	0.17851	0.02149
0.55	0.556	0.09482	0.00518
0.50	0.500	0.0	0.0

If $\lambda \leq 0.5$, then any positive amount of study is unremunerative.

Appendix D

SOME RESULTS OF OPTIMAL STRATEGIES TO ACHIEVE A PASSING GRADE

Let us suppose a student faces a test consisting of N questions, each with two alternative answers. He knows the test will be scored using an admissible scoring system S, but that the only thing that matters is that his total score exceeds a certain "passing threshold" T. Suppose all the questions "look alike" to him, in the sense that on each question he feels there is probability $p \geq \frac{1}{2}$ that one alternative is correct, and probability $1 - p$ that the other is correct. Assume also that the questions are independent (in the stochastic sense). Then the student will perceive that his chance of ascribing the higher probability to the correct alternative on exactly K out of N questions is exactly

$$\binom{N}{K} p^K (1 - p)^{N-K} . \tag{D.1}$$

If he makes the same response ($(r, 1 - r)$, $r > \frac{1}{2}$) on each question, then the value (V) of his score, if he ascribes the higher probability to the correct alternative on K out of N questions, will be

$$V(K, r) = KS(r) + (N - K)S(1 - r) . \tag{D.2}$$

If $r > \frac{1}{2}$, $S(r) > S(1 - r)$. Therefore, if $K_1 > K_2$, then $V(K_1, r) > V(K_2, r)$. Now let us define $K^*(r)$ as follows:

$$K^*(r) = \frac{T - NS(1 - r)}{S(r) - S(1 - r)} . \tag{D.3}$$

$K^*(r)$ is not necessarily an integer. By virtue of the above equation, however, if K is an integer such that $K > K^*(r)$, then

$$KS(r) + (N - K)S(1 - r) > T . \tag{D.4}$$

Therefore, the student will maximize his probability of "passing the test" (i.e., getting a score greater than T) if he selects that r which minimizes $K^*(r)$. I assert that the optimum value of r (which we shall call r^*) is that r which satisfies the equation,

$$r^*S(r^*) + (1 - r^*)S(1 - r^*) = \frac{T}{N} . \tag{D.5}$$

By substituting (D.5) in (D.3) we see that

$$\frac{K^*(r^*)}{N} = r^* . \tag{D.6}$$

Now consider some $r \neq r^*$, $r > \frac{1}{2}$. We will show that $K^*(r) > K^*(r^*)$, thus proving that r^* is an optimal response. By definition of what an admissible scoring system is, we know that

$$\frac{K^*(r^*)}{N} S(r^*) + \left(1 - \frac{K^*(r^*)}{N}\right)S(1 - r^*) > \frac{K^*(r^*)}{N} S(r) + \left(1 - \frac{K^*(r^*)}{N}\right)S(1 - r) . \tag{D.7}$$

From this, and (D.3), we deduce

$$K^*(r)S(r) + (N - K^*(r))S(1 - r) = T > K^*(r^*)S(r) + (N - K^*(r^*))S(1 - r) . \tag{D.8}$$

Thus

$$K^*(r)[S(r) - S(1 - r)] > K^*(r^*)[S(r) - S(1 - r)]$$

$$K^*(r) > K^*(r^*). \tag{D.9}$$

The fact that the solution to (D.5) is an optimal response in this setting is a striking illustration of the fact that nonlinear utility for score may destroy the admissible property of a scoring system, for

Eq. (D.5) does not depend upon the student's subjective probability p at all!

It is interesting to note, by the way, that the optimal total test strategy may not involve making the same response on all questions, even when the student's subjective probabilities on all questions are the same. For example, if T = S(1) and he is completely uninformed ($p = \frac{1}{2}$) on all questions, then he will secure a 50 percent chance of passing by making a (1,0) response on one question and a $(\frac{1}{2}, \frac{1}{2})$ response on all the rest. This is manifestly a better chance than he can secure by any strategy that calls for the same (r, 1 - r) response on every question.

BEST COPY AVAILABLE

REFERENCES

1. Brown, T. A., and H. Shuford, Jr., *Quantifying Uncertainty Into Numerical Probabilities for the Reporting of Intelligence*, The Rand Corporation, R-1185-ARPA, 1973.

2. Grayson, C., *Decisions Under Uncertainty: Drilling Decisions by Oil and Gas Operators*, Harvard Business School, Division of Research, Boston, Mass., 1950.

3. Savage, L. J., "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, Vol. 66, 1971, pp. 783-801.

4. Sibley, W. L., *A Prototype Computer Program for Interactive Computer-Administered Admissible Probability Measurement*, The Rand Corporation, R-1258-ARPA, 1973.

5. Shannon, C. E., and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill., 1949.

6. Shuford, H., Jr., A. Albert, and H. E. Massengill, "Admissible Probability Measurement Procedures," *Psychometrika*, Vol. 31, 1966, pp. 125-145.

7. Brown, T. A., *Probabilistic Forecasts and Reproducing Scoring Systems*, The Rand Corporation, RM-6299-ARPA, June 1970.